Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling*

MATHIEU LAGRANGE AND SYLVAIN MARCHAND
(lagrange@labri.fr) (sylvain.marchand@labri.fr)

LaBRI, Université Bordeaux 1, F-33405 Talence Cedex, France

AND

JEAN-BERNARD RAULT
(jeanbernard.rault@francetelecom.com)

France Telecom R&D, F-35512 Cesson Sevigné cedex, France

Within the context of sinusoidal modeling, a new method for the interpolation of sinusoidal components is proposed. It is shown that autoregressive modeling of the amplitude and frequency parameters of these components allows us to interpolate missing audio data realistically, especially in the case of musical modulations such as vibrato or tremolo. The problem of phase discontinuity at the gap boundaries is also addressed. Finally, an original algorithm for the interpolation of a missing region of a whole set of sinusoids is presented. Objective and subjective tests show that the quality is improved significantly compared to common sinusoidal and temporal interpolation techniques of missing audio data.

0 INTRODUCTION

The sinusoidal model [1], [2] provides a high-quality representation of pseudostationary sounds. Therefore this model is used widely for many musical audio processing purposes such as musical sound processing [3]–[5] and audio coding [6], [7]. Parameters of the sinusoidal model are extracted from the original sound in a frame-based manner, and a sound that is close to the original one can be synthesized from the extracted parameters.

The problem of missing information about sinusoids can occur at both sides of the sinusoidal analysis and synthesis procedure. During the analysis some gaps in the original signal may have been introduced by another module, for example, a module of detection and removal of clicks or transients. During the synthesis, sinusoidal parameters may not be available. For example, in a stream-based audio coding application, some frame packets may be unavailable at the time they are needed for the synthesis. In both cases, information about the sinusoids is available before and after the gap and can be exploited to interpolate the evolution of the partials within the missing region.

Let a gap start at frame index n_1 and end at frame index n_2, corrupting a set of sinusoids $\boldsymbol{S}$. The aim of the algorithm described in this paper is to interpolate $\boldsymbol{S}$ during the gap. As shown in Fig. 1, the set $\boldsymbol{B}$ represents sinusoids existing before the gap and ending at frame n_1. The set $\boldsymbol{A}$ represents sinusoids existing after the gap and beginning at frame n_2. Only the sinusoids of these two sets will be considered for the interpolation of the gap.

The block diagram in Fig. 1(a) describes the four-step algorithm used to interpolate the missing region. The predicted frequencies and amplitudes in the missing region are computed for each sinusoid of the two sets [Fig. 1(b)]. According to these predicted parameter sets $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{A}}$, some sinusoids of $\boldsymbol{B}$ are matched to sinusoids of $\boldsymbol{A}$. These matched sinusoids then become sinusoids with a missing region [dashed lines in Fig. 1(c)]. This missing region is interpolated using the predicted parameters of the two matched sinusoids. Next, unmatched sinusoids (terminating or beginning with open dots) are extrapolated in the missing region according to their predicted parameters using a specific technique. The interpolated set of sinusoids $\hat{\boldsymbol{S}}$ is plotted in Fig. 1(c).

The remainder of this paper is organized as follows. The sinusoidal model and the limitation of existing interpolation methods are presented in Section 1. The use of autoregressive (AR) modeling for the prediction of the amplitude and frequency parameters of a sinusoid in a missing region is presented in Section 2. Section 3 describes the matching of sinusoids from both sides of the missing region and introduces the use of the predicted parameters to enhance the matching of modulated sinusoids. Next an original method for interpolating the missing parameters of a partial is introduced in Section 4 and is followed by objective and subjective evaluations of this interpolation method. The extrapolation of unmatched sinusoids is presented in Section 5. Finally an algorithm for the interpolation of a whole set of sinusoids in a missing region that makes use of these concepts is compared in Section 6 to known sinusoidal and temporal techniques.

*Manuscript received 2004 December 6; revised 2005 July 28.

1 SINUSOIDAL MODELING

Sinusoidal modeling aims at representing a sound signal as a sum of sinusoids of given amplitudes, frequencies, and phases. For stationary pseudoperiodic sounds these amplitudes and frequencies evolve slowly and continuously with time, controlling a set of pseudosinusoidal oscillators commonly called partials. (This term will be preferred to sinusoid during the remainder of this paper.) The audio signal s can be calculated from the additive parameters using Eqs. (1) and (2),

$$s(t) = \sum_{p=1}^{P} A_p(t) \quad \cos\left[\phi_p(t)\right] \tag{1}$$

$$\phi_p(t) = \phi_p(0) + 2\pi \int_0^t f_p(u)\, du \tag{2}$$

where P is the number of partials and the functions f_p, A_p, and ϕ_p are the instantaneous frequency, amplitude, and phase of the pth partial, respectively. The P triplets (f_p, A_p, ϕ_p) are the parameters of the additive model and represent points in the frequency–amplitude plane at time t.

Although potential applications are numerous, few people have paid attention to the interpolation issue. Quatieri and Danisewicz [8] propose an algorithm to interpolate overlapping harmonics for the purpose of separating two speech signals. The amplitude is interpolated linearly, and cubic interpolation is used for the phase. The frequency can be found by the differentiation of the cubic phase polynomial. Although this strategy was originally designed for intraframe parameter interpolation for synthesis purposes [1], this method shows good results for gaps of lengths from 20 to 100 ms during stationary regions of speech sounds. Later on Maher [9] proposed an algorithm to interpolate a whole set of sinusoids with an approximation of missing audio data based on the same principles.

This interpolation method based on a polynomial interpolation of the parameters of the partials preserves the harmonic relation among partials together with the envelope of the sound. Yet modulations of the parameters of the partials are not taken into account. For example, the frequency of a partial having natural vibrato is a sinusoid in the time–frequency plane of about 4-Hz frequency. Since the phase polynomial is cubic, the resulting interpolation of the frequency is a quadratic polynomial. A sinusoid is approximated correctly by a quadratic polynomial for less than a quarter of a period. The use of such an interpolation scheme for frequency and phase parameters is limited to segments up to 60 ms. Similarly, if we want to handle natural tremolo, the use of linear interpolation is limited to segments of up to 20 ms.

According to Bregman [10] these modulations should be considered, because such modulations play an important role in sound perception: “Small fluctuations in frequency occur naturally in the human voice and in musical

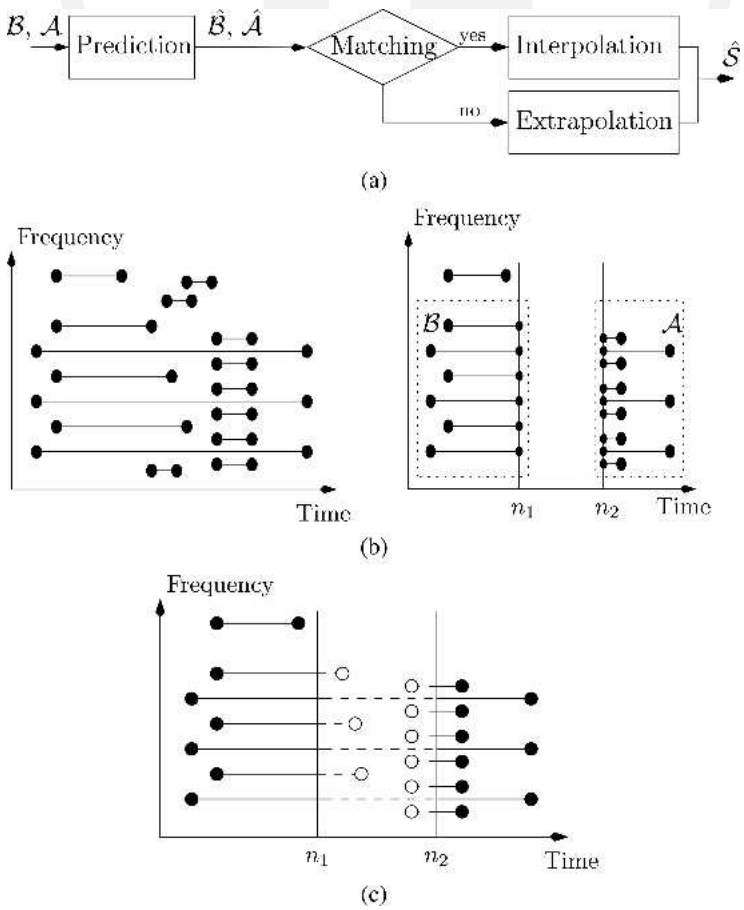

Fig. 1. (a) Block diagram of proposed interpolation algorithm. (b) Left, original set of sinusoids ***S*** and right, sets of sinusoids (***A*** and ***B***) used during the interpolation process. (c) Interpolated set of sinusoids ***S***.

instruments. The fluctuations are not often very large, ranging from less than 1 percent for a clarinet tone to about 1 percent for a voice trying to hold a steady pitch, with larger excursions of as much as 20 percent for the vibrato of a singer. Even the smaller amounts of frequency fluctuation can have potent effects on the perceptual grouping of the components' harmonics." Although Bregman only talks about frequency modulations, amplitude modulations are important too.

Comments of the experts who performed the listening test presented in Section 4.5 confirmed this assertion. The missing region interpolated using the polynomial scheme was perceived as many simple tones and not as a complex one. As a result, the interpolated part was perceived as artificial. To achieve a more natural interpolation, one needs an interpolation method able to preserve these modulations of the frequency and the amplitude of partials in the missing region.

Linear prediction has proven successful for digital audio restoration [11]. Given the AR modeling of parts of the signal before and after the degradation, linearly predicted extrapolations can be added to interpolate the degraded part of the signal (see [12], [13] for further details). Considering that evolutions of the amplitude and frequency parameters of the partials are time signals too—although with a much lower sampling rate—a similar strategy can be used for the extrapolation and interpolation of amplitudes and frequencies of the partials.

2 PREDICTING EVOLUTION OF THE PARTIALS

Let P_i and P_j denote partials of the $\boldsymbol{B}$ and $\boldsymbol{A}$ sets, respectively,

$$P_i = \{P_i(n), n = n_1 - l_i + 1, \cdots, n_1\} \tag{3}$$

$$P_j = \{P_j(n), n = n_2, \cdots, n_2 + l_j - 1\} \tag{4}$$

$$P_k(n) = (f_k(n), A_k(n), \phi_k(n)), \quad \text{for all } k \tag{5}$$

where l_i and l_j are the lengths of P_i and P_j respectively, and $P_k(n)$ is the triplet of instantaneous parameters of the partial P_k at frame n.

Let $\hat{P}_i$ and $\hat{P}_j$ denote the predicted amplitude and frequency of the partials of the $\boldsymbol{B}$ and $\boldsymbol{A}$ sets during the missing region,

$$\hat{P}_i = \{\hat{P}_i(n_1 + k), k = 1, \cdots, n_2 - n_1 - 1\} \tag{6}$$

$$\hat{P}_j = \{\hat{P}_j(n_2 - k'), k' = 1, \cdots, n_2 - n_1 - 1\} \tag{7}$$

$$\hat{P}_k(n) = (\hat{f}_k(n), \hat{A}_k(n)), \quad \text{for all } k \tag{8}$$

where $\hat{P}_k(n)$ is a couple of instantaneous predicted parameters since the phase will not be predicted, but deduced from the frequency. These parameters should be computed using a relevant method, chosen according to the characteristics of the evolutions of the amplitude and frequency parameters. These evolutions in the time–frequency and time–amplitude planes can be constant, increasing or decreasing exponentially (portamento in the time–frequency plane) or sinusoidal (vibrato in the time–frequency plane and tremolo in the time–amplitude plane).

We propose that these parameters can be modeled by an AR model [14], [15], and linear prediction (LP) is used in order to predict the parameters of the partials in the missing region. In LP the current sample $x(n)$ is approximated by a linear combination of past samples of the input signal,

$$\hat{x}(n) = \sum_{h=1}^{\kappa} a_K(h)x(n-h) \tag{9}$$

where K is the order of the LP model. We are then looking for a vector a_K that minimizes the power of the prediction error,

$$E = \sum_{n=1}^{N} [x(n) - \hat{x}(n)]^2 \tag{10}$$

Supposing that a vector a_K minimizing the power of the prediction error of the frequencies of P_i is known, the frequencies and amplitudes of $\hat{P}_i$ are computed by infinite impulse response filtering of the frequencies and amplitudes of P_i [see Fig. 5(b)]. The same strategy is applied to the partial P_j, except that the two extrapolations are done backward [see Fig. 5(c)].

As it will be demonstrated, this extrapolation scheme is able to preserve the modulations of the parameters of the partials in the missing region. However, the predictions of the frequencies of the partials of a harmonic source are computed separately. The proposed prediction scheme also preserves harmonicity provided that the partials of $\boldsymbol{B}$ and $\boldsymbol{A}$ are estimated correctly. Let us consider a set of partials with harmonically related frequencies. The fundamental is denoted by P_0 and the harmonics by P_r, with $r > 0$. The frequencies of the harmonics verify

$$F_r(n) = (r + 1)F_0(n). \tag{11}$$

To predict the evolution of the frequencies of these partials, we consider LP coefficients for each harmonic $a_K^r(h)$ computed using $F_r(n)$ as observations. Because of Eq. (11) and the scale invariance of LP coefficients [16], we have $a_K^r = a_K^0$. Thus the harmonicity constraint is preserved,

$$\hat{F}_r(n) = (r + 1)\,\hat{F}_0(n). \tag{12}$$

2.1 Linear Prediction Methods

The challenge in linear prediction is to choose a well-suited method to minimize the error E, given N past samples—considered as observations—and the model order K. In this section three methods are described out of many: the autocorrelation method, the covariance method, and the Burg method. Only the method retained is detailed so that it can be implemented easily; the reader is invited to refer to [17], [16] for a complete description of the others. The choice among these three methods is driven by specific constraints: only few observed samples are available and the estimated LP coefficients have to be suitable for extrapolation.

The autocorrelation method minimizes the forward prediction error power on an infinite support. In practice the signal is finite. Samples of the $x(n)$ process that are not observed are then set to zero, and observed samples are

windowed in order to minimize the discontinuity at the boundaries. As a consequence, this method requires $N >> K$ to be effective.

Alternatively, the LP coefficients can be estimated on a finite support with the covariance method. This method minimizes the forward prediction error power on a finite support. Since no zeroing of the data is necessary, this method is a good candidate for coefficient estimation of a process having few observed samples. Unfortunately this method should be avoided for data extrapolation because it can lead to filters that are not minimum phase, that is, the estimated poles are not guaranteed to lie within the unit circle.

Let $e_k^{\mathrm{f}}(n)$ and $e_k^{\mathrm{b}}(n)$ denote, respectively, the forward and backward prediction errors at an intermediate order k,

$$e_k^{\mathrm{f}}(n) = x(n) + \sum_{h=1}^{k} a_k(h)x(n-h) \tag{13}$$

$$e_k^{\mathrm{b}}(n) = x(n-k) + \sum_{h=1}^{k} a_k(h)x(n-k+h). \tag{14}$$

The Burg method minimizes the average of the forward and backward error power on a finite support in a recursive manner. That is, to obtain a_k we minimize

$$\epsilon_k = \frac{1}{2}\left(\epsilon_k^{\mathrm{f}} + \epsilon_k^{\mathrm{p}}\right) \tag{15}$$

where

$$\epsilon_k^{\mathrm{f}} = \frac{1}{(N-k)} \sum_{n=k}^{N-1} |e_k^{\mathrm{f}}(n)|^2 \tag{16}$$

$$\epsilon_k^{\mathrm{b}} = \frac{1}{(N-k)} \sum_{n=0}^{N-k-1} |e_k^{\mathrm{b}}(n)|^2 \tag{17}$$

and

$$a_k(h) = \begin{cases} a_{k-1}(h) + r_k a_{k-1}(k-h), & h = 1, 2, \cdots, k-1 \\ r_k, & h = k \end{cases} \tag{18}$$

r_k being the reflection coefficient. By substituting Eq. (18) in Eqs. (16) and (17) we find a recursive expression for the forward and backward errors,

$$e_k^{\mathrm{f}}(n) = e_{k-1}^{\mathrm{f}}(n) + r_k e_{k-1}^{\mathrm{b}}(n-1) \tag{19}$$

$$e_k^{\mathrm{b}}(n) = e_{k-1}^{\mathrm{b}}(n-1) + r_k e_{k-1}^{\mathrm{f}}(n) \tag{20}$$

where

$$e_0^{\mathrm{f}}(n) = e_0^{\mathrm{b}}(n) = x(n). \tag{21}$$

To find r_k we differentiate the kth prediction error power with respect to r_k, and by setting the derivative to zero we obtain

$$r_k = \frac{-2\sum_{n=k}^{N-1} e_{k-1}^{\mathrm{f}}(n) e_{k-1}^{\mathrm{b}}(n-1)}{\sum_{n=k}^{N-1} |e_{k-1}^{\mathrm{f}}(n)|^2 + |e_{k-1}^{\mathrm{b}}(n-1)|^2}. \tag{22}$$

The minimum-phase property is ensured because the expression of r_k is of the form $r_k = 2xy/(|x|^2 + |y|^2)$, where x and y are vectors. Using the Schwarz inequality, it is verified that r_k has a magnitude lower than 1.

With the Burg method the minimization is done on a finite support and the joint minimization of the forward and backward errors leads to a stable filter. This method is then suitable for data extrapolation with few observed samples. The following algorithm computes the vector a of LP coefficients at order K using the Burg method,

$e_{\mathrm{f}} \leftarrow x$
$e_{\mathrm{b}} \leftarrow x$
$a \leftarrow 1$

for k from 1 to K do

$e_{\mathrm{f}p} \leftarrow e_{\mathrm{f}}$ without its first element
$e_{\mathrm{b}p} \leftarrow e_{\mathrm{b}}$ without its last element
$r_k \leftarrow -2e_{\mathrm{b}p} \cdot e_{\mathrm{f}p}/(e_{\mathrm{b}p} \cdot e_{\mathrm{b}p} + e_{\mathrm{f}p} \cdot e_{\mathrm{f}p})$
$e_{\mathrm{f}} \leftarrow e_{\mathrm{f}p} + r_k e_{\mathrm{b}p}$
$e_{\mathrm{b}} \leftarrow e_{\mathrm{b}p} + r_k e_{\mathrm{f}p}$
$a \leftarrow (a(0), a(1), \cdots, a(k), 0) + r_k(0, a(k), a(k-1), \cdots, a(0))$

end for.

2.2 Linear Prediction Parameters

The number of observed samples used to estimate the LP coefficients has to be large enough to be able to extract the signal periodicity, and short enough not to be too constrained by the past evolution. In our system the short-term analysis module uses a sliding time–frequency transform with a hop size of 360 samples on sound signals sampled at CD quality (44.1 kHz). This means that the frequency and amplitude trajectories are sampled at ≈120 Hz. Since we want to handle natural vibrato with a frequency of about 4 Hz, we need at least 30 samples to get the period of the vibrato. For frequency and amplitude evolutions, since we want to model exponentially increasing or decreasing evolutions (portamento) and sinusoidal evolutions (vibrato, tremolo), the order of the LP model should not be below 2. Most modulations are more complex than the sinusoidal behavior of vibrato or tremolo, thus the order should be set at a higher value.

The LP coefficients used to compute the predicted parameters $\hat{P}_i$ and $\hat{P}_j$ are estimated using the Burg method. This method jointly minimizes the forward and backward prediction errors defined by Eqs. (16) and (17). As a consequence the number of observed samples must be at least twice the model order. In the experiments presented here, N is chosen as the minimum value between 40 and l_i or l_j, respectively, and the model order m is set to the integer value closest to $N/2$.

3 MATCHING PARTIALS FROM BOTH SIDES OF THE MISSING REGION

The first step to interpolate corrupted sinusoidal data in the missing region is to decide which partial of $\boldsymbol{B}$ should be linked to which partial of $\boldsymbol{A}$ to form a unique partial. The problem of matching partials from both sides of the missing region is shown in Fig. 2. The time interval is so long that the evolution of the partials within the missing

region has to be taken into account to achieve a good match. We propose that this decision step can be done using predicted information ($\hat{P}_i$ and $\hat{P}_j$) computed using the method introduced in the previous section.

This issue is quite similar to the partial tracking problem, but with a much longer time interval between elements to be linked. First a straightforward adaptation of the partial tracking algorithm proposed in [1] is discussed. It will be used in Section 6 for comparison purposes. Couples of partials (P_i, P_j) such that the distance between the last frequency of P_i and the first frequency of P_j is below a given threshold Δ_f are matched,

$$|f_i(n_1) - f_j(n_2)| < \Delta_f \tag{23}$$

where $f_i(n_1)$ is the last frequency of P_i and $f_j(n_2)$ is the first frequency of P_j, and Δ_f is a threshold parameter in hertz. Yet if the spectrum is changing within the gap interval, this approach may be unsatisfactory, as explained in [9] and shown in Fig. 4(a).

Considering that the parameters of the partials have a predictable evolution is useful to match the partials of the two ***B*** and ***A*** sets more reliably. Unfortunately, considering a simple Euclidean distance between the two predictions in frequency or amplitude may lead to difficulties. If the two predictions vary a lot, the thresholding procedure should be more tolerant than if the two predictions are nearly constant (see Fig. 3). To cope with this problem, a Euclidean distance between the two predictions normalized by the sum of the standard deviation of the two predictions is used to decide whether or not partials from both sides of the missing region should be matched.

Let $d_f(P_i, P_j)$ denote the normalized Euclidean distance between the predicted frequencies $\hat{f}_i$ and $\hat{f}_j$,

$$d_f(P_i, P_j) = \sqrt{\frac{\sum_{n=n_1+1}^{n_2-1} (\hat{f}_i(n) - \hat{f}_j(n))^2}{n_2 - n_1 - 1}}. \tag{24}$$

The normalized Euclidean distance $d_A(P_i, P_j)$ between the predicted amplitude is defined similarly. Each couple of partials (P_i, P_j) such that $d_f(P_i, P_j)$ is below a given threshold Δ_f is a candidate for matching.

Next these candidates are considered in increasing d_f distance order. The candidate partials are effectively matched if two criteria involving predicted frequencies and predicted amplitudes are satisfied. These criteria are defined as

$$\frac{d_f(P_i,P_j)}{1 + \sigma(\hat{f}_i) + \sigma(\hat{f}_j)} < T_f \tag{25}$$

$$\frac{d_A(P_i, P_j)}{1 + \sigma(\hat{A}_i) + \sigma(\hat{A}_j)} < T_a \tag{26}$$

where $\sigma(x)$ is the standard deviation of the vector x, and T_f and T_a are threshold parameters in frequency and amplitude.

If these criteria are met for a couple (P_i, P_j), the two partials of the couple are merged in a unique partial P_m, and each couple where P_i or P_j appears is removed from the sorted list. The missing region of the resulting partial $\hat{P}_m$ is interpolated using the method described in the next section. This process iterates until no satisfactory couple remains. Using this algorithm, the matching is performed even in modulated cases (see Fig. 4) without spurious link in stationary cases [see Fig. 3(b)]. Finally unmatched partials are extrapolated in the missing region using an algorithm described in Section 5.

4 INTERPOLATING THE MISSING INFORMATION WITHIN A PARTIAL

Let a couple (P_i, P_j) be represented as a unique partial P_m. The interpolated frequency and amplitude parameters

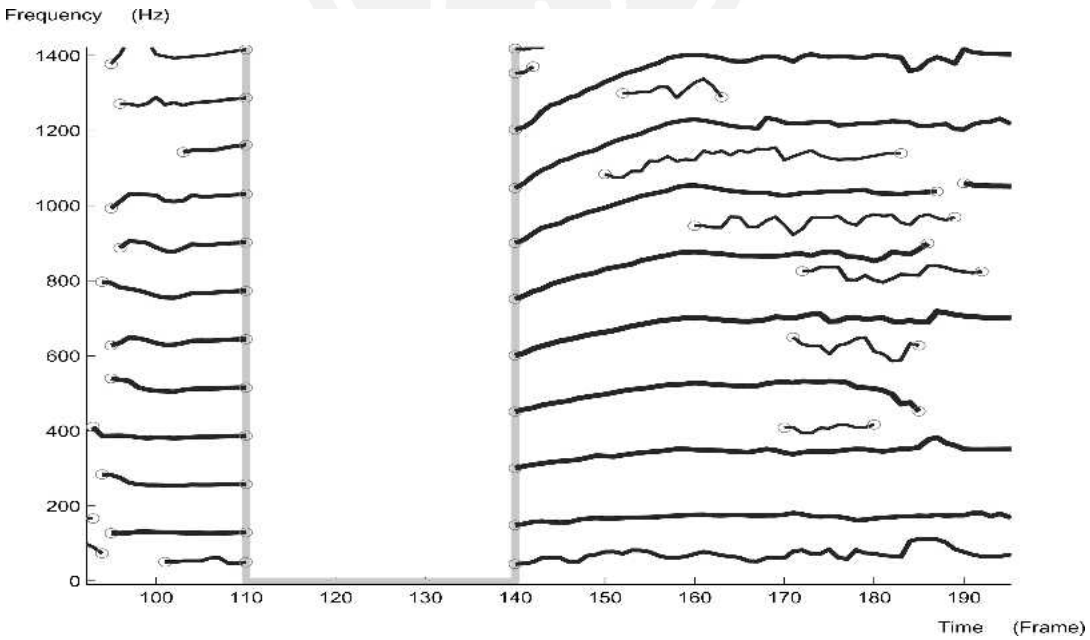

Fig. 2. Matching partials from both sides of missing region.

of $\hat{P}_m$, starting at $n_1 + 1$ and ending at $n_2 - 1$, are computed by mixing the predicted frequency and amplitude parameters $\hat{P}_i$ and $\hat{P}_j$. The phase continuity at the boundaries of the missing region is then ensured by a method described at the end of this section.

4.1 Frequency Interpolation

To compute $\hat{f}_m(n)$ given the two predicted frequencies $\hat{f}_i(n)$ and $\hat{f}_j(n)$, a crossfading is carried out by multiplying $\hat{f}_i(n)$ by a window function w and $\hat{f}_j(n)$ by $1 - w$,

$$\hat{f}_m(n) = w\left(\frac{n-n_1}{n_2-n_1}\right)\hat{f}_i(n) + \left[1 - w\left(\frac{n-n_1}{n_2-n_1}\right)\right]\hat{f}_j(n). \tag{27}$$

As can be seen in Fig. 5, the forward prediction is of better quality than the backward one, since here P_i is longer than P_j. In general the window function $w(t)$ used to crossfade the two predictions should then be asymmetric in order to favor the prediction done with the largest data set. The symmetric cosine window computed using Eq. (28) is equal to 0.5 in the middle of the missing region,

$$c(t) = \frac{1 + \cos[\pi(1+t)]}{2}. \tag{28}$$

The symmetric crossfading done using this window function is relevant only if the two partials P_i and P_j have the

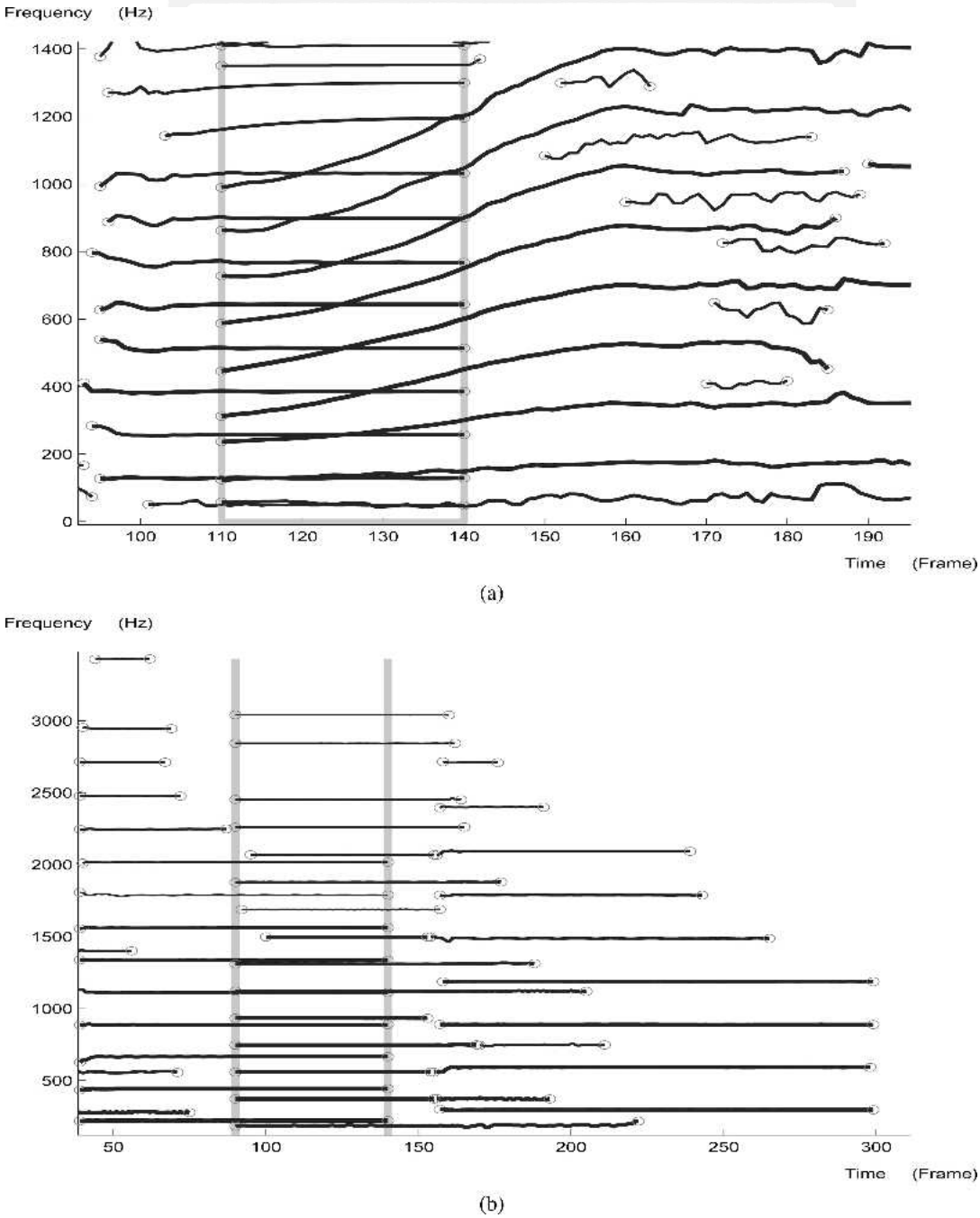

Fig. 3. Predictions of partials from both sides of missing region. (a) Trombone tone with glissando. (b) Transition between two piano tones.

same length. If P_i is three times longer than P_j, the window should reach the 0.5 value at 3/4 of the missing region (see Fig. 6). As a consequence, the window function must fulfill the following constraint:

$$w\left(\frac{l_i}{l_i + l_j}\right) = \frac{1}{2}. \tag{29}$$

We propose that such an asymmetric crossfading can be done using an asymmetric factor;

$$r(x, y) = \frac{\log(1/2)}{\log\{c[x/(x+y)]\}} \tag{30}$$

where log is the Neperian logarithm.

This factor is computed according to l_i and l_j, the respective lengths of P_i and P_j. The asymmetric window function is then

$$w(t) = \begin{cases} c(t)^{r(l_i,l_j)}, & l_i > l_j \\ 1 - [1 - c(t)]^{r(l_j,l_i)}, & \text{otherwise} \end{cases} \tag{31}$$

with $t \in [0, 1]$.

4.2 Amplitude Interpolation

The amplitude of a partial is often much more modulated than its frequency, as in speech signals. Even if mi-

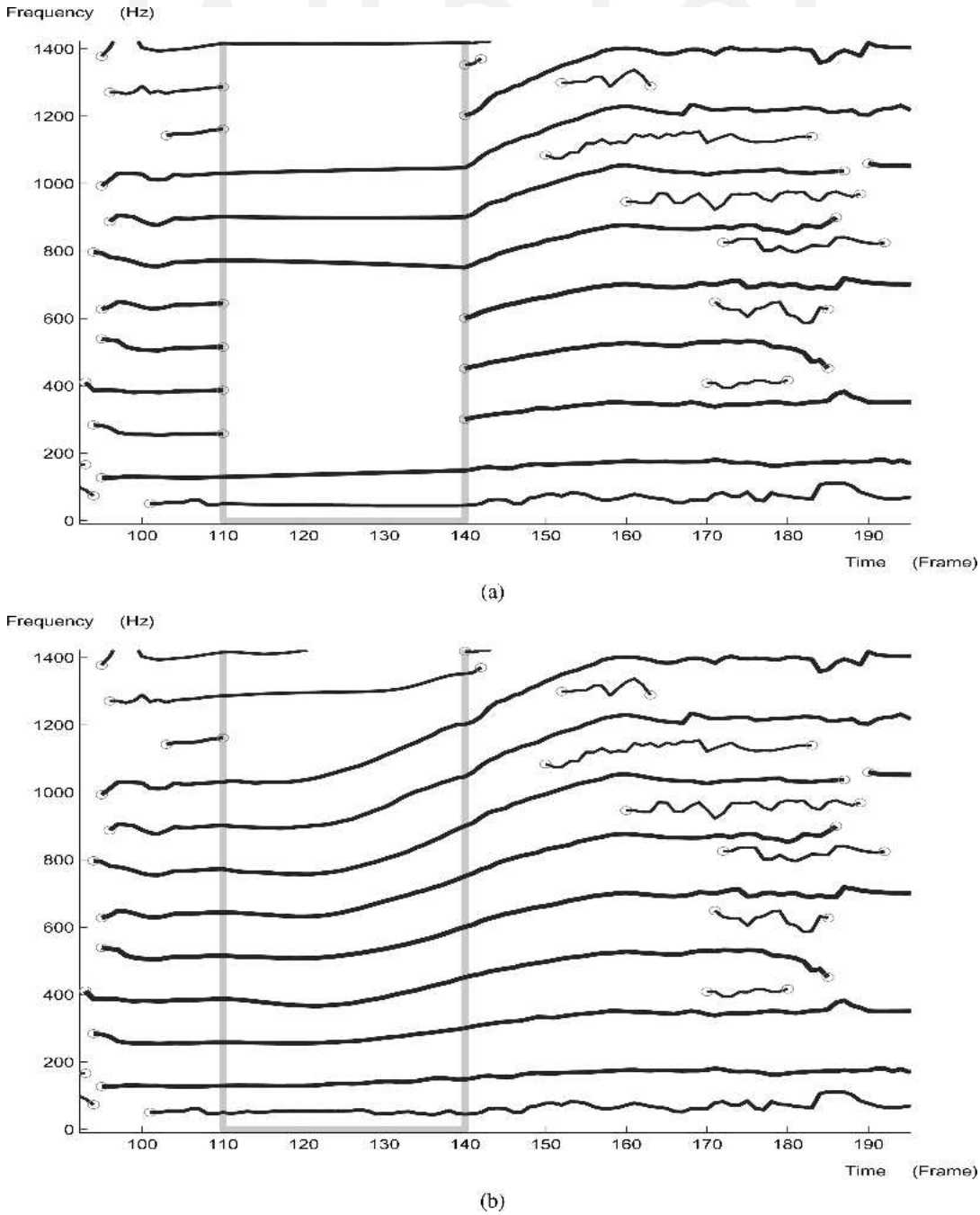

Fig. 4. Results of matching process. (a) Reference method. (b) Proposed method.

cromodulations of the amplitude parameter are preserved, the long-term prediction is not satisfactory.

Before the crossfade the amplitude prediction of the partial P_i is constrained to end at a given amplitude equal to the mean amplitude of the partial P_j computed from frame n_2 to frame $\min(n_2 + M, n_2 + l_j - 1)$. The parameter M should be chosen so as to get an energy estimate of the beginning of partial P_j. In the configuration presented in Section 2.2, M is set to 30. Such a constraint is fulfilled by adding to the predicted amplitude $\hat{A}_i$ an increment $\delta_i(n)$ defined as

$$\delta_i(n) = \frac{n - n_1}{n_2 - n_1}\left[\frac{\sum_{\tau=0}^{\min(M, l_j - 1)} \hat{A}_j(n_2 + \tau)}{\min(M, l_j - 1) + 1} - \hat{A}_i(n_2)\right]. \quad (32)$$

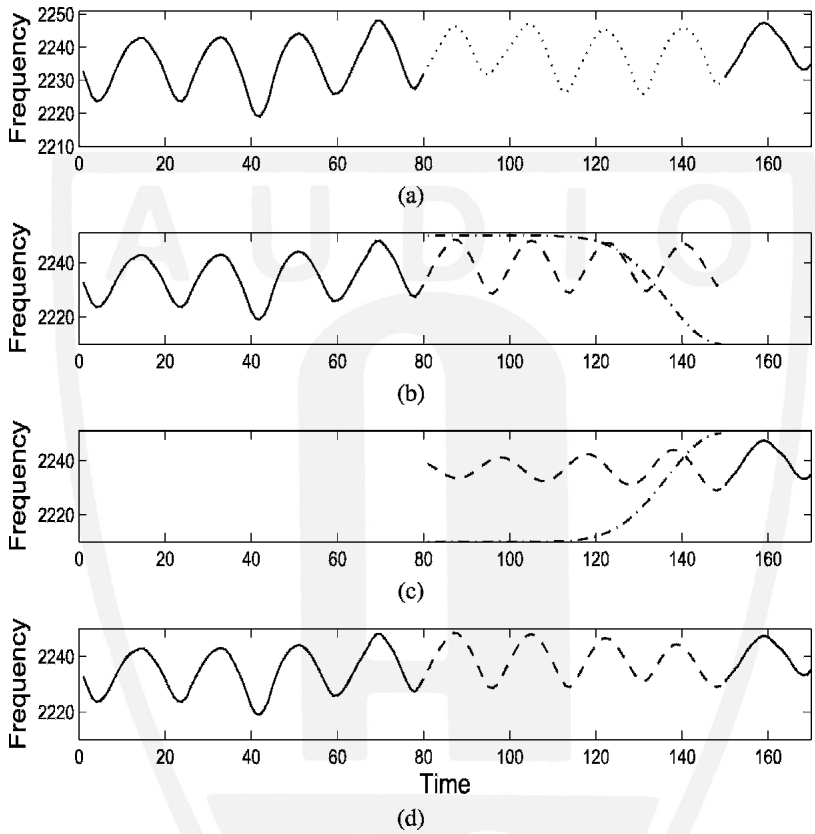

Fig. 5. Interpolating frequencies of a partial of a saxophone with vibrato using AR modeling. Forward prediction; – – – backward prediction with LP formalism; – · – predictions crossfaded using an asymmetric window favoring the more reliable predicted samples (those of the forward prediction in this case). (a) Frequencies represented by dots are unavailable.

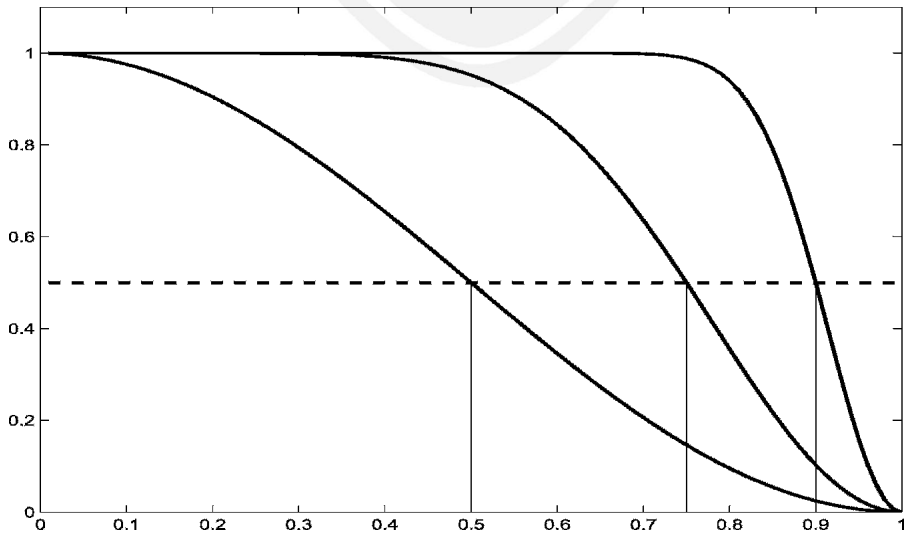

Fig. 6. Three crossfading windows computed using Eq. (31). From left to right, windows are computed with $l_i/l_j \in \{2/1, 3/1, 9/1\}$.

The same strategy is applied to $\hat{A}_j$ by adding an increment $\delta_j(n)$ computed as follows:

$$\delta_j(n) = \frac{n_2 - n}{n_2 - n_1}\left[\frac{\sum_{\tau=0}^{\min(M,l_i-1)} A_i(n_1-\tau)}{\min(M, l_i - 1) + 1} - \hat{A}_j(n_1)\right]. \quad (33)$$

The corrected amplitudes are then asymmetrically crossfaded to provide the interpolated amplitude,

$$\hat{A}_m(n) = w\left(\frac{n-n_1}{n_2-n_1}\right)[\hat{A}_i(n) + \delta_i(n)] + \left[1 - w\left(\frac{n-n_1}{n_2-n_1}\right)\right][\hat{A}_j(n) + \delta_j(n)]. \quad (34)$$

4.3 Phase Interpolation

Using the interpolation strategy described in [8], the phase of a partial P_m is interpolated using a maximally smooth cubic polynomial having four constraints at the boundaries $f_m(n_1)$, $\phi_m(n_1)$ and $f_m(n_2)$, $\phi_m(n_2)$. The interpolated frequencies are then obtained by phase differentiation.

Inversely we propose to integrate the predicted frequencies $\hat{f}_m$ using the trapezoidal method. A phase increment ρ defined below is added to each phase in the missing region to ensure phase continuity at boundaries. Let us denote by $\varphi(n)$ the unwrapped phase at frame n, [$\phi(n) = \varphi(n) \bmod 2\pi$]. The subscript m is omitted for convenience.

In a first approximation the missing phases may be computed from

$$\tilde{\varphi}(n_1 + 1) = \phi(n_1) + \pi T[f(n_1) + \hat{f}(n_1 + 1)] \quad (35)$$

$$\tilde{\varphi}(n) = \tilde{\varphi}(n_1 + 1) + \pi T \sum_{\tau=n_1+1}^{n} [\hat{f}(\tau - 1) + \hat{f}(\tau)] \quad (36)$$

where $n \in [n_1 + 2, n_2]$ and T is the hop size in seconds. However, a phase discontinuity may occur at the end of the missing region: $\tilde{\phi}(n_2) \neq \phi(n_2)$. Let e_ϕ denote the error of the phase extrapolation at n_2,

$$e_\phi = \tilde{\phi}(n_2) - \phi(n_2). \quad (37)$$

We satisfy the continuity constraint of phase by "spreading" the error through the whole missing region. The interpolated phases are then computed during the missing region as follows:

$$\hat{\varphi}(n) = \tilde{\varphi}(n) + \frac{n - n_1}{n_2 - n_1}\rho \quad (38)$$

where $n \in [n_1 + 1, n_2]$ and ρ is chosen to ensure the continuity constraint at the end boundary: $\hat{\phi}(n_2) - \phi(n_2) = 0$. Since $\phi(n_2)$ is a known 2π modulus, the number of solutions for ρ is infinite. The smallest one is retained,

$$\rho = \begin{cases} e_\phi + 2\pi, & |e_\phi| < -\pi \\ e_\phi - 2\pi, & |\hat{e}_\phi| > \pi \\ e_\phi, & \text{otherwise.} \end{cases} \quad (39)$$

Given the predicted amplitudes and frequencies of the partials from both sides of the missing region, we are able to interpolate reliably the missing region of a partial. The capability of this new interpolation scheme will be evaluated in the remainder of this section, where a synthetized version of the interpolated sinusoidal representation is compared to the signal synthetized from the original sinusoidal representation.

4.4 Objective Evaluation

We simulate a missing region in the sinusoidal representation $\boldsymbol{S}$ by deleting parameters of the partials existing before and after the missing region. The other partials are left as they are, as illustrated in Fig. 7. Missing parameters of partials are then interpolated during the missing region using the polynomial or the LP-based interpolation scheme.

In all the experiments reported here the interpolation scheme described in [1] is used for the intraframe interpolation of the parameters of the partials. The amplitude is interpolated linearly and phases are computed using a maximally smooth cubic polynomial. The reconstruction signal-to-noise ratio (R-SNR) is used to evaluate the performance of the algorithm tested,

$$\text{R-SNR} = 10\log_{10}\left(\frac{\sum_{m=0}^{M-1}[x(m) - \hat{x}(m)]^2}{\sum_{m=0}^{M-1} x^2(m)}\right) \quad (40)$$

where $x(m)$ is the original temporal signal and $\hat{x}(m)$ the synthesized signal of the sinusoidal representation interpolated using one of the two tested interpolation strategies. For every gap size the result plotted in Fig. 8 is the mean R-SNR for every position of the gap.

The LP-based interpolation is designed for musical modulation management (vibrato, tremolo) and therefore performs better for the saxophone tone or the vibraphone tone and performs as well as the polynomial method in the stationary case, such as for the harpsichord tone.

4.5 Subjective Evaluation

The two methods are compared by a subjective test performed at France Telecom R&D with ten experts in audio processing. Four audio signals were used: a saxophone tone, a vibraphone tone, a soprano female voice, and an orchestra piece. It tests the interpolation for gap sizes from 80 to 820 ms. For every gap size and audio file, the experts were asked to listen to the original set of partials synthesized as an explicit reference signal. After this first listening, they were asked to note four versions, one with no interpolation performed, one with interpolation performed using the polynomial approach, one with the interpolation performed using the LP approach, and the original set of partials synthesized as a hidden reference.

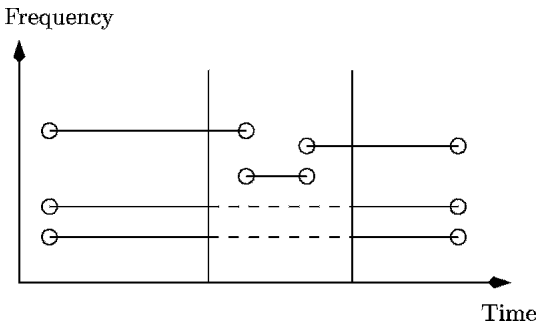

Fig. 7. Simulating a missing region.

They were asked to note these four versions using the 100-point Mushra scale. The marks obtained by the two interpolated versions are plotted in Fig. 9.

As can be seen in Fig. 9(a), a high-quality interpolation of monophonic signals having vibrato (saxophone tone) or tremolo (vibraphone tone) for missing region sizes close to 1 s is achieved. The audio signals having more complex modulations, such as the singing voice or the orchestra piece, are harder to interpolate, but the LP-based method is a significant improvement [Fig. 9(b)].

5 EXTRAPOLATION OF UNMATCHED PARTIALS

Considering that the matching between partials of $\boldsymbol{B}$ and $\boldsymbol{A}$ is done correctly, unmatched partials of $\boldsymbol{B}$ belong to a note decaying in the missing region and unmatched sinusoids of $\boldsymbol{A}$ belong to a note that started in the missing region (see Fig. 1).

Let $l_{\boldsymbol{B}}$ be the maximum length of the extrapolation of unmatched partials of $\boldsymbol{B}$ and $l_{\boldsymbol{A}}$ the maximum length of the extrapolation of unmatched partials of $\boldsymbol{A}$. The extrapolation of unmatched partials P_i or P_j is done according to the predicted parameters $\hat{P}_i$ or $\hat{P}_j$. The predicted frequencies $\hat{f}_i$ or $\hat{f}_j$ are used as is, and the extrapolated phases are computed using Eq. (36).

In general, the amplitudes of the partials have a predictable behavior during the ending of the note (sustain or decay). The predicted amplitude of the unmatched partial of $\boldsymbol{B}$ can then be used safely to detect at which frame the partial should end. The extrapolated amplitude $\tilde{A}_i(n)$ is then the predicted amplitude $\hat{A}_i(n)$ faded as follows:

$$\tilde{A}_i(n) = \hat{A}_i(n) - \gamma_i(n) \tag{41}$$

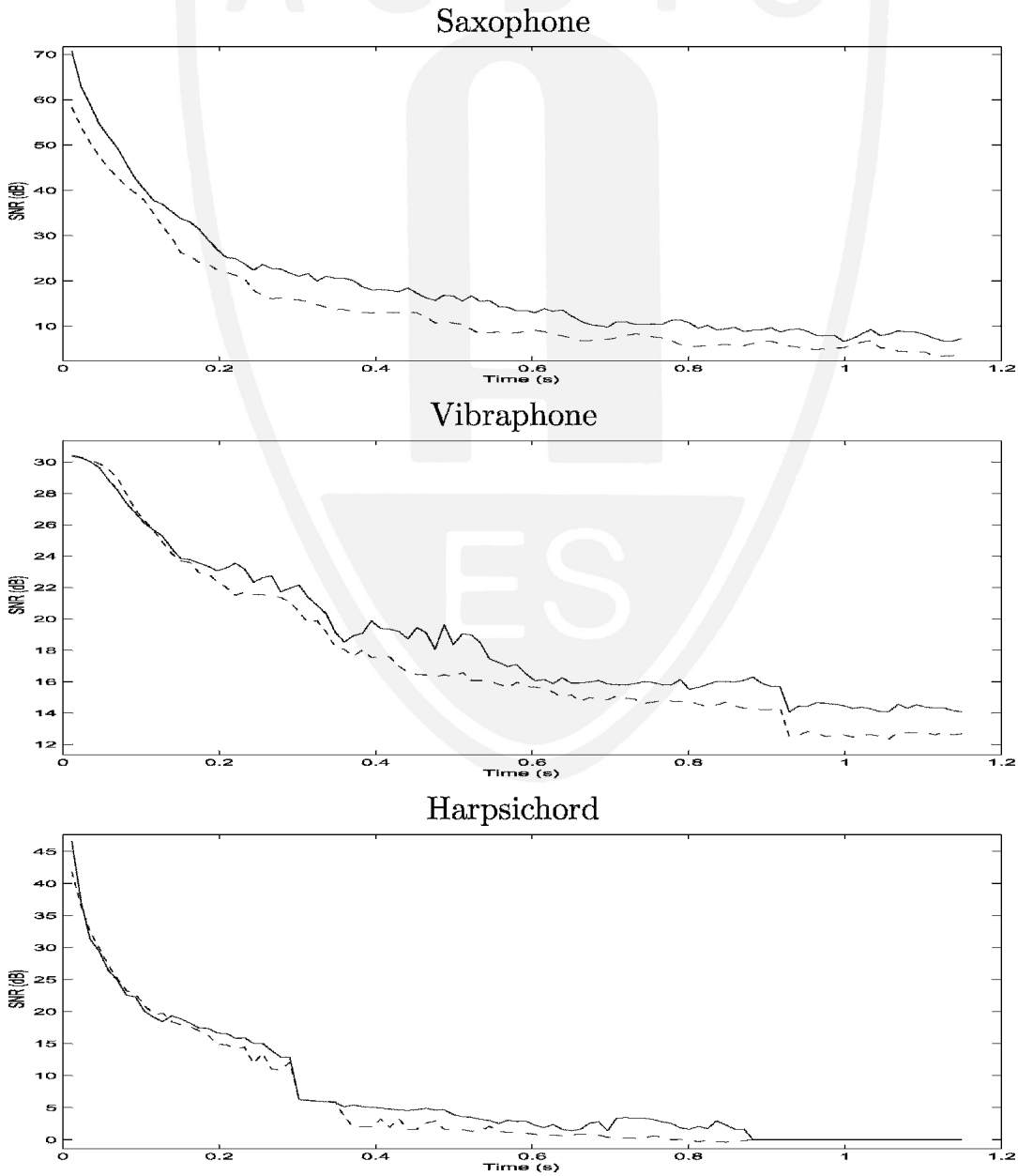

Fig. 8. Objective comparison of LP-based interpolation (———) and polynomial interpolation (– – –) on three sound signals. (a) Saxophone tone with vibrato. (b) Vibraphone. (c) Harpsichord.

with

$$\gamma_i(n) = \frac{n - n_1}{l_B} \max\left[\hat{A}_i(n_1 + l_B), 0\right]. \quad (42)$$

If the extrapolated amplitude $\tilde{A}_i(n)$ becomes negative at a frame $n < n_1 + l_B$, the extrapolated partial $\tilde{P}_i$ ends at frame $n - 1$. As a consequence the partial may end before $n_1 + l_B$, as shown in Fig. 1.

On the other hand, the amplitude of the partials during an abrupt onset cannot be deduced from the amplitude of the partial during the sustain part of the note. To at least simulate an onset, all unmatched partials of $\boldsymbol{A}$ should begin at the same frame index $n_2 - l_A$. The extrapolated amplitude $\tilde{A}_j(n)$ is then the predicted amplitude $\hat{A}_j(n)$ faded as follows:

$$\tilde{A}_j(n) = \hat{A}_j(n) - \gamma_j(n) \quad (43)$$

with

$$\gamma_j(n) = \frac{n_2 - n}{l_A} \hat{A}_j(n_2 - l_A). \quad (44)$$

The extrapolated partial $\tilde{P}_j$ starts at the smaller frame index $n \geq n_2 - l_A$ so that $\tilde{A}_j(n + k) > 0$, for all $k \geq 0$.

The parameters l_B and l_A should be chosen according to the targeted application. For interpolating the sinusoidal data lost due to a transmission error, the maximum gap size allowed is generally small due to the limited data buffering capability of the decoder. In this approach the extrapolation should be parameterized to be tolerant to mismatch that occurred during the matching step of the algorithm. This can be done by setting $l_B = l_A = n_2 - n_1 - 1$ to ensure a fade in or out of unmatched sinusoids. During the sinusoidal analysis step or with a digital data restoration application, however, some extra information about the spectral content can be used to estimate the frame index where the unmatched partials of $\boldsymbol{A}$ should start. The parameter l_A can be set to an onset index estimate for every gap occurring.

6 INTERPOLATION OF MISSING AUDIO DATA

This section compares three methods of prediction for missing audio data using subjective listening tests with the same protocol as the one used in Section 4.5. The temporal method uses 2000 temporal samples from both sides of the

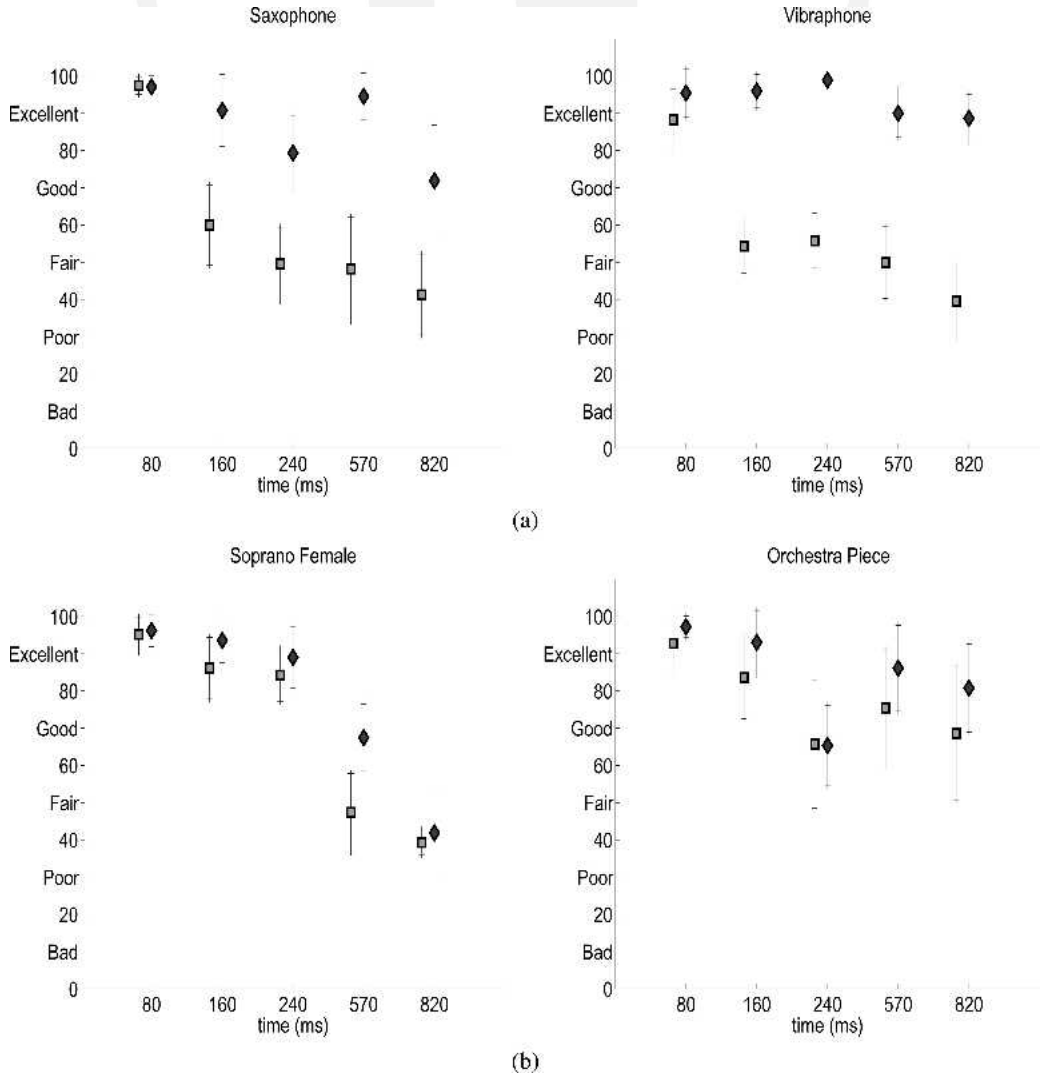

Fig. 9. Results of listening tests comparing polynomial-based method (■) and LP-based method (♦) for five gap sizes. Symbols—means of votes; lines—confidence intervals for each method.

region to estimate two sets of 1000 LP coefficients using the Burg method. The predictions obtained by filtering are crossfaded using the window computed using Eq. (28).

The two other methods are based on sinusoidal modeling. First, two sets of partials ($\boldsymbol{B}$ and $\boldsymbol{A}$) are extracted using the sinusoidal analysis technique described in [15]. The interpolated set of partials $\hat{\boldsymbol{S}}$ is computed using one of the two sinusoidal schemes and synthesized.

With the polynomial method the matching of partials is done according to Eq. (23) with $\Delta_f = 40$ Hz. The missing phases and frequencies are computed using the maximally smooth cubic phase polynomial whereas the amplitude is interpolated linearly. Extrapolated parameters of unmatched partials are computed using the algorithm detailed in Section 5, considering constant frequencies and amplitudes as predicted parameters.

With the proposed method the matching is done using the algorithm described in Section 3 with $T_f = 0.5$ and $T_a = 0.1$. Interpolation of the missing parameters of the partials uses the method described in Section 4, and the extrapolated parameters of the unmatched partials are computed using the algorithm detailed in Section 5 with $l_{\boldsymbol{B}} = n_2 - n_1 - 1$ and $l_{\boldsymbol{A}} = (n_2 - n_1 - 1)/2$.

Five audio signals are used: a violin tone with vibrato, a piano tone, an orchestra piece, a gong tone, and the recording of two female soprano voices. The gap can be at a sustained or at a transitional segment of the sound.

LP-based temporal interpolation has proven successful for the interpolation of up to thousands of samples from CD-quality audio signals without audible distortion [13], [18]. The interpolation quality of longer gaps depends on the characteristics of the signal. If it consists of stationary partials like in the piano tone, the attenuation phenomenon is lightly pronounced [see Fig. 10 (a), left]. Yet if the interpolated signal has roughly the same number of partials—around ten—with vibrato, the attenuation is very pronounced [see Fig. 10 (a), right]. This attenuation problem explains why the marks obtained by this method range from 30 to 50 when the parameters of the partials are modulated (see Fig. 11). The sinusoidal model can be used to cope with this attenuation problem [see Fig. 10(b), (c)].

The sinusoidal interpolation scheme based on polynomial interpolation outperforms the temporal method for gap sizes up to 320 ms (see Fig. 11). In counterpart, all kinds of modulations disappear. This effect is perceived by the listeners as a "freezing" of the sound throughout the interpolated region. For larger gaps linear interpolation gives an artificial interpolation, rated poorly by the listeners. The rating can be even worse than the one obtained by the temporal method. This is the case for the interpolation of a 820-ms gap of the violin tone (see Fig. 11).

The proposed method keeps the advantages of the two previous methods while avoiding some of their disadvantages. Use of a sinusoidal model avoids the problem of attenuation as long as long-gap interpolation can be achieved. In addition AR modeling of the parameters of the partials is useful to preserve the modulations important to perception. The gong tone and the two soprano voices have partials with small-range modulations. The violin tone with vibrato has a larger range of frequency modulations. For all these sounds the ratings go from 90 to 70 in a regular decay for gap sizes from 320 to 820 ms. The soprano voices can even be interpolated during 1.6 s with a *good* mark. The partials extracted from the orchestra piece have complex modulations because they represent harmonics of several notes and noise. The prediction capability is then lower than in the previous cases, but a *fair* quality can be achieved for gap sizes up to 450 ms.

If the gap occurs during a transition, some important information is lost and the quality of interpolation is lower, (see Fig. 12). In this case the temporal scheme seems to be better appreciated, probably because of the attenuation effect that simulates a fade in/fade out centered at the middle of the gap. Concerning sinusoidal interpolation schemes, the quality is improved by the use of the matching algorithm presented in Section 3, which is useful to avoid a mismatch of partials of different tones.

7 CONCLUSION

In this paper an enhanced method is proposed for the interpolation of audio signals based on linear prediction in sinusoidal modeling. It is shown that AR modeling of the parameters of the partials allows those partials to be interpolated reliably. Partials having simple modulations such as vibrato or tremolo allow high-quality interpolation for gap sizes up to 1 s. More complex modulations are harder to interpolate, but the proposed method shows a significant improvement over the polynomial method. Since these modulations are important to perception [10], the sinusoidal interpolation of missing audio data is more realistic. The listening tests showed that the proposed method provides fair interpolation for complex polyphonic signals for gap sizes up to 450 ms and good interpolation for monophonic modulated tones for gap sizes up to 1600 ms.

8 REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 34, pp. 744–754 (1986).

[2] J. O. Smith and X. Serra, "An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *Proc. Int. Computer Music Conf. (ICMC)* (Computer Music Assoc., San Francisco, CA, 1987).

[3] K. R. Fitz and L. Haken, "Sinusoidal Modeling and Manipulation Using Lemur," *Computer Music J.,* vol. 20, no. 4, pp. 44–59 (Winter 1996).

[4] X. Serra, "Musical Sound Modeling with Sinusoids plus Noise," in *Musical Signal Processing,* Studies on New Music Research ser. (Swets & Zeitlinger, Lisse, The Netherlands, 1997) pp. 91–122.

[5] S. Marchand and R. Strandh, "InSpect and ReSpect: Spectral Modeling, Analysis and Real-Time Synthesis Software Tools for Researchers and Composers," in *Proc. Int. Computer Music Conf. (ICMC)* (International Computer Music Assoc., Beijing China, 1999, Oct.), pp. 341–344.

[6] H. Purnhagen and N. Meine, "HILN—The MPEG-4 Parametric Audio Coding Tools," in *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS 2000),* vol. 3 (2000 May), pp. 201–204.

[7] B. den Brinker, E. Schuijers, and W. Oomen, "Parametric Coding for High-Quality Audio," presented at the 112th Convention of the Audio Engineering Society, *J. Audio. Eng. Soc. (Abstracts),* vol. 50, p. 510 (2002 June), convention paper 5554.

[8] T. F. Quatieri and R. G. Danisewicz, "An Approach to Co-channel Talker Interference Suppression Using a Sinusoidal Model for Speech," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 38, pp. 56–69 (1990 Jan).

[9] R. C. Maher, "A Method for Extrapolation of Miss-

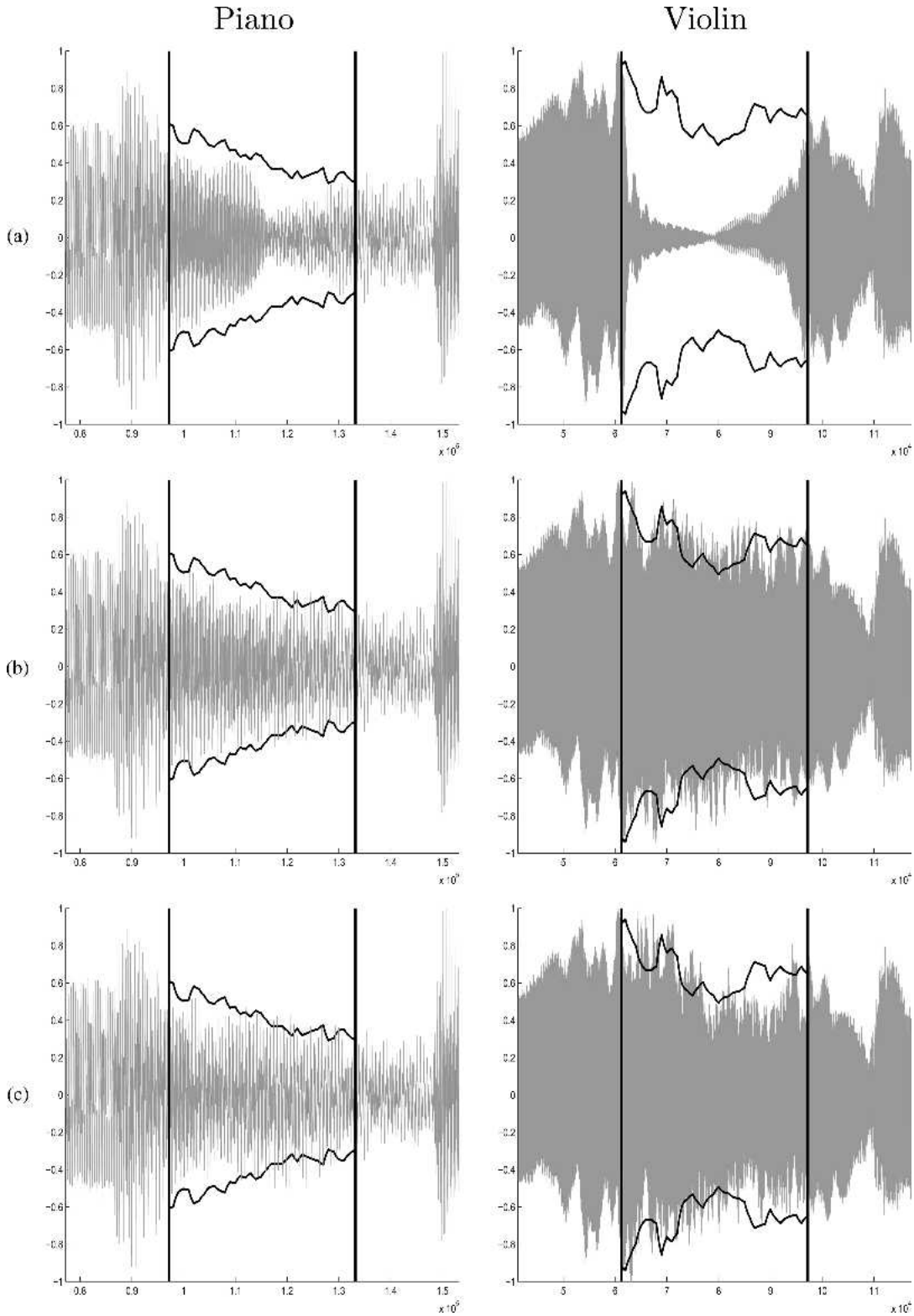

Fig. 10. Temporal representations of piano tone and violin tone with vibrato, interpolated using three methods tested during 820 ms. Two vertical lines fix boundaries of missing region; two symmetric lines inside this region approximate envelope of original sound. (a) Temporal interpolation. (b) Polynomial-based sinusoidal interpolation. (c) LP-based sinusoidal interpolation.

ing Digital Audio Data," *J. Audio Eng. Soc.* (*Engineering Reports*), vol. 42, pp. 350–357 (1994 May).

[10] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA, 1990).

[11] A. J. E. M. Janssen, R. N. J. Veldhius, and L. B. Vries, "Adaptive Interpolation of Discrete-Time Signals that Can Be Modeled as Autoregressive Processes," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 34, pp. 317–330 (1986).

[12] W. Etter, "Restoration of a Discrete-Time Signal Segment by Interpolation Based on the Left-Sided and

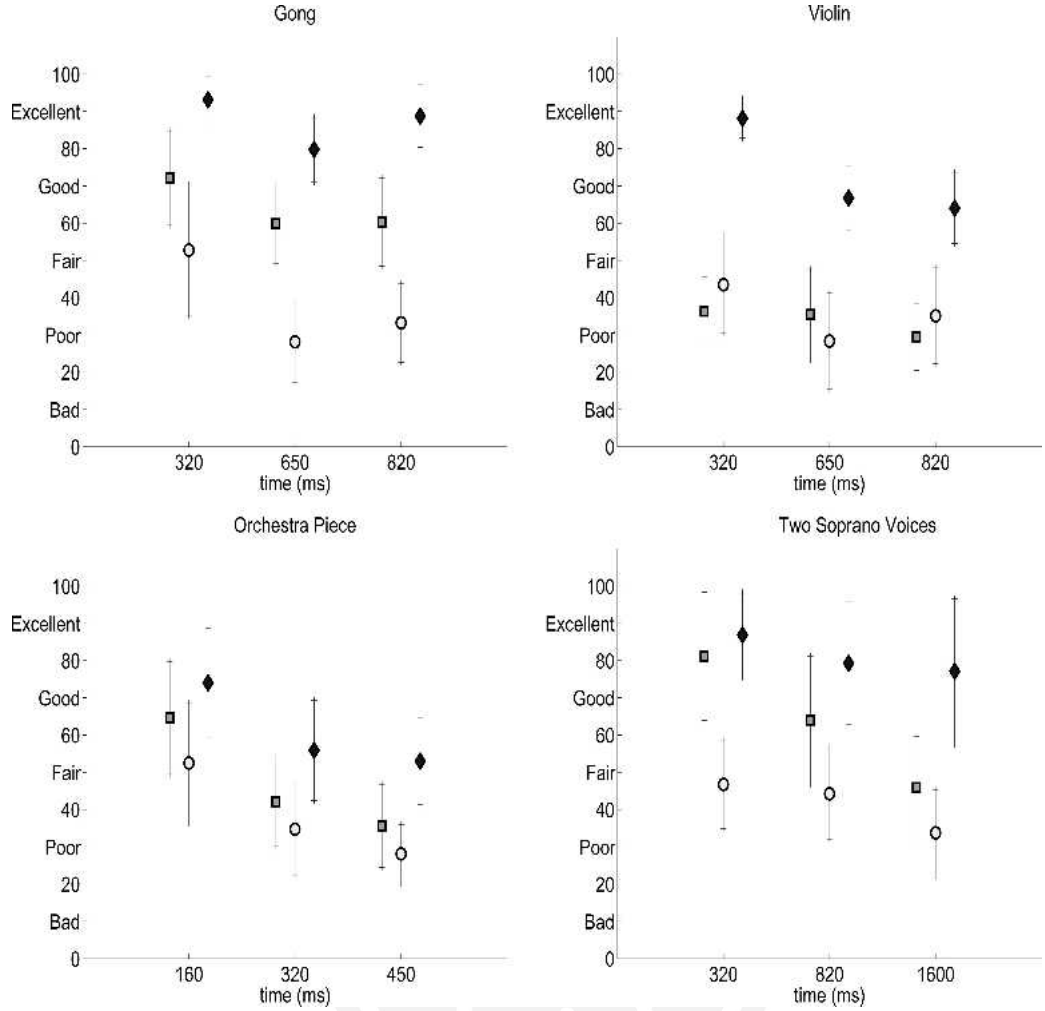

Fig. 11. Results of listening tests comparing polynomial-based method (■), LP-based method (♦), and temporal method (○) for three gap sizes. Symbols—means of votes, lines—confidence intervals for each method.

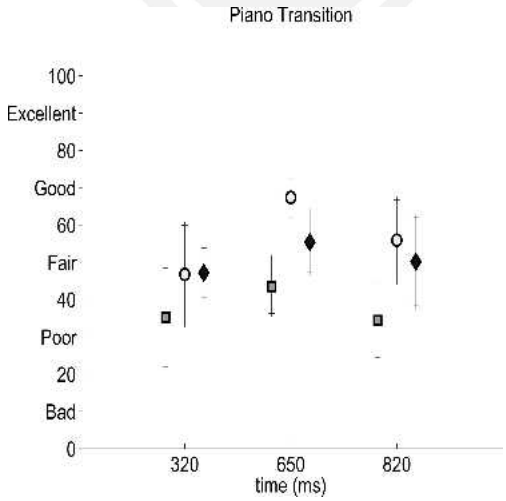

Fig. 12. Results of the listening tests comparing polynomial-based method (■), LP-based method (♦), and temporal method (○) for three gap sizes on a transitional segment of piano. Symbols—means of votes, lines—confidence intervals for each method.

Right-Sided Autoregressive Parameters," *IEEE Trans. Acoust., Speech, Signal Process,* vol. 44, pp. 1124–1135 (1996).

[13] I. Kauppinen, J. Kauppinen, and P. Saarinen, "A Method for Long Extrapolation of Audio Signals," *J. Audio Eng. Soc.,* vol. 49, pp. 1167–1180 (2001 Dec.).

[14] M. Lagrange, S. Marchand, M. Raspaud, and J. B. Rault, "Enhanced Partial Tracking Using Linear Prediction," in *Proc. Digital Audio Effects (DAFx) Conf.* (Queen Mary, University of London, 2003 Sept.), pp. 141–146.

[15] M. Lagrange, S. Marchand, and J. B. Rault, "Using Linear Prediction to Enhance the Tracking of Partials," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* vol. 4 (2004 May), pp. 241–244.

[16] S. M. Kay, "Autoregressive Spectral Estimation: Methods," in *Modern Spectral Estimation,* Signal Processing ser. (Prentice-Hall, Englewood Cliffs, NJ, 1988), pp. 228–231.

[17] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE,* vol. 63, pp. 561–580 (1975 Nov.).

[18] I. Kauppinen and K. Roth, "Audio Signal Extrapolation—Theory and Applications," in *Proc. Digital Audio Effects (DAFx) Conf.* (University of the Federal Armed Forces, Hamburg, Germany, 2002 Sept.), pp. 105–110.

THE AUTHORS

M. Lagrange

S. Marchand

J.-B. Rault

Mathieu Lagrange was born in Caen, France, in 1978. He studied computer science at the University of Rennes 1, France, where he obtained his master's degree in 2000. He received a postgraduate diploma with a focus on spectral sound synthesis from the University of Bordeaux 1, Talence, France.

Dr. Lagrange carried out research on sound analysis and coding at the France Telecom Laboratories in partnership with the LaBRI (computer science laboratory), University of Bordeaux 1, there he received a Ph.D. degree in 2004. He is particularly involved in spectral sound analysis, audio restoration, and auditory scene analysis. He is a member of SCRIME (Studio de Création et de Recherche en Informatique et Musique Electroacoustique) at the University.

●

Sylvain Marchand was born in Pessac near Bordeaux, France, in 1972. He studied computer science at the University of Bordeaux 1, Talence, France. He obtained his master's degree in 1995 and a postgraduate diploma in algorithmics the following year. In the meantime he carried out research in computer music and sound modeling. He received a Ph.D. degree in 2000.

Dr. Marchand was appointed associate professor at the LaBRI (computer science laboratory), University of Bordeaux 1, in 2001. He is particularly involved in spectral sound analysis, transformation, and synthesis. He is a member of SCRIME (Studio de Création et de Recherche en Informatique et Musique Electroacoustique) at the University.

●

Jean-Bernard Rault received a Ph.D. degree in signal processing and telecommunications from the University of Rennes, France, in 1987.

Dr. Rault then joined the CCETT in Rennes, France, to collaborate on the European project Eureka 147 (DAB) in the area of digital audio compression. From 1990 to 1992 he spent two years at Thomson-LER, there he was involved in the Multicarrier Digital Modulation studies. Since 1993 he has been a France Telecom representative with ISO/MPEG and participates in the development of the MPEG Audio coding standards. He has also been involved in several European projects (MoMuSys, Cinenet, Nadib, Song, Ardor) to contribute to audio-related work packages.