Estimation of Frequency for AM/FM Models Using the Phase Vocoder Framework

Michaël Betser, Patrice Collen, Gaël Richard, *Senior Member, IEEE*, and Bertrand David, *Member, IEEE*

***Abstract*—This paper proposes an extension of the applicability of phase-vocoder-based frequency estimators for generalized sinusoidal models, which include phase and amplitude modulations. A first approach, called phase corrected vocoder (PCV), takes into account the modification of the Fourier phases resulting from these modulations. Another approach is based on an adaptation of the principles of the time-frequency reassignment and is referred to as the reassigned vocoder (RV). The robustness of the estimation against noise is studied, both theoretically and experimentally, and the performance is assessed in comparison with two state-of-the-art algorithms: an unmodified version of the reassignment method and a quadratically interpolated fast Fourier transform method (QIFFT).**

***Index Terms*—AM/FM model, frequency estimation, phase vocoder.**

I. Introduction

FOR several decades, sinusoidal parameters estimation has remained one of the most popular topics in the field of signal processing and numerous approaches have been proposed. Some of the most popular estimators are based on Fourier analysis [1]–[7], on nonlinear least squares analysis [8], [9] or on subspace methods, such as ESPRIT [10] and MUSIC [11].

The signal $s(t)$ studied by these methods is often expressed in the form of a complex sinusoid perturbed by noise

$$s(t) \triangleq x(t) + n(t) \triangleq \mathrm{e}^{L(t)+j\Phi(t)} + n(t) \qquad \text{(I.1)}$$

where the log-amplitude $L(t)$ and the phase $\Phi(t)$ are both real functions of the discrete time t, and $n(t)$ is a white Gaussian noise ($n \in \mathbb{C}$). Herein after, the case for multiple sines will be considered as an extension of the single-component case, as long as they are sufficiently resolved in frequency.

L and Φ are usually assumed to be analytical functions and the model (I.1) can thus be locally approximated by a polynomial model which can be considered as a truncated Taylor expansion in the neighborhood of t:

$$L_K(t+\tau) \triangleq \sum_{k=0}^{K} l^{(k)}(t)\frac{\tau^k}{k!}, \quad \Phi_L(t+\tau) \triangleq \sum_{k=0}^{L} \phi^{(k)}(t)\frac{\tau^k}{k!} \qquad \text{(I.2)}$$

Widely studied models include the following:

- **M01**. $K = 0$, $L = 1$: the simplest model where constant-amplitude and constant-frequency sines are considered;
- **M11**. $K = 1$,$L = 1$: the first-order AM model (also known as exponential sinusoidal model or ESM) which considers exponentially modulated amplitudes;
- **M02**. $K = 0$, $L = 2$L: the first-order FM model (also known as chirp model) that considers linearly frequency-modulated sinusoids;
- **M12**. $K = 1$, $L = 2$: a more general, first-order AM/FM, model where both frequency-modulated and amplitude-modulated sinusoids are considered.

The AM model has recently been actively investigated, in order to describe transients [12], as for example percussive sounds in music, speech attacks, and free vibrating systems [13] such as plucked strings. The FM model has also been widely used in speech and music, in analysis–transformation–synthesis schemes [14], and to describe glissando or transition between phonemes. Other applications have been investigated in the fields of radar and sonar [15], [16], and seismology [17]. Both models have been found useful for coding purposes [18]. Speech and music signals are by nature nonstationary. Thus, amplitude and frequency are time-varying, and these variations are sometimes too important to be neglected over the length of analysis. Moreover, amplitude and frequency modulations occur simultaneously, as in speech attacks and decays. This suggests that frequency estimation methods for the AM/FM model will prove to be useful in audio applications.

For estimating the M01-model parameters, the maximum-likelihood (ML) frequency estimation scheme leads to take the maximum of the periodogram [1], [19], as the ML-estimate. When more complex models are considered, however, the problem moves to a multidimensional optimization problem [20]–[22]. This optimization usually involves time-consuming iterative computation [23]. It is therefore interesting to develop simpler estimators, though suboptimal. This is the standpoint of this paper, which is applied in the context of Fourier based frequency estimation.

As mentioned before, a wide number of such estimators exists, mostly based on the simplest sinusoidal model M01. For FM models (M02,M12), good efficiency can be achieved by means of the reassignment method [2]. A recent approach

Manuscript received October 24, 2006; revised May 8, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Steven M. Kay.

M. Betser and P. Collen are with France Telecom R&D, 35512 Cesson-Sévigné, France (e-mail: michael.betser@orangeft-group.com; patrice.collen@orangeft-group.com).

G. Richard and B. David are with the Département TSI, GET-ENST, 75014 Paris, France (e-mail: gael.richard@enst.fr; bertrand.david@enst.fr).

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/TSP.2007.906768

[the quadratically interpolated fast Fourier transform method (QIFFT)] based on an adaptation of an earlier spectrum interpolation method proposes to consider the more general AM/FM signal model in the case of Gaussian analysis windows [14], [24]. All these estimators use an initial computation of a time-frequency transform on a discrete grid for one or two (in the case of the reassignment) different analysis windows and further combine the grid values to derive the parameter estimates.

From the early times of the phase-vocoder [25], the currently used frequency estimation method consists of deriving a phase difference between two successive values in the same channel of the transform. Starting from some considerations relative to this particular frequency estimator in the AM/FM context, this paper will in particular introduce a new approach that considers phase differences *across* different channels.

The paper is organized as follows. The principles of the classical phase vocoder are summarized in Section II. The extension of the phase vocoder to the FM model are presented in Section III along with two new estimators, namely the phase-corrected vocoder (PCV) and the reassigned vocoder (RV). Then, it is shown in Section IV that the reassigned vocoder can be directly applied to the first-order AM/FM model. The theoretical influence of noise on the phase vocoder is studied in Section V. Experimental comparisons with existing estimators are given in Section VI, and finally some conclusions are drawn in Section VII.

II. Phase Vocoder

The term "vocoder," derived from "voice coder," originally refers to a speech analyzer and synthesizer. The phase vocoder uses a polar representation of the short-time spectrum [25]. Instantaneous frequency estimation is at the heart of the method and is computed as a discrete derivative of the phase.[1] This analysis/synthesis method is known for its ability to analyze [26], modify [27]–[29], and resynthesize [30] a sound, and for its use as an electronic musical instrument [31].

A. Phase Vocoder Analysis Framework

The phase vocoder analysis framework is based on the uniform-rate short-time Fourier transform (STFT). Hereafter, the zero-phased (or centered) form of the Fourier transform (FT)[2] is used.

$$X(t_m,\omega_k;h) \triangleq \sum_{n=-(N-1)/2}^{(N-1)/2} x(\tau_n+t_m)h(\tau_n)\mathrm{e}^{-j\tau_n\omega_k} \quad \text{(II.1)}$$

where h is the analysis window, N is the sample size of the Fourier transform, F_s is the sampling frequency, k is the frequency bin, and $\tau_n \triangleq n/F_s$ is the time in seconds of the corresponding sample number n. The time corresponding to the center of the STFT window is noted $t_m \triangleq m/F_s$. Finally, $\omega_k \triangleq 2\pi k F_s/N$ is the frequency of the bin k.

The signal processing based on the vocoder framework relies on the definition of the local model $x(t_M+\tau)=x_{\mathrm{loc}}(\tau)$. This model is assumed to be valid in the neighborhood of t_M, and in particular, on the analysis interval centered in t_M, with a length W. This interval usually includes a few overlapping frames ($W > N$).

B. Basic Phase Vocoder Scheme

The basic version of the phase vocoder hypothesizes the linear phase model M01, i.e., locally constant frequency and amplitude:

$$x_{\mathrm{loc}}(\tau) \triangleq A\mathrm{e}^{j(\alpha_M+\beta\tau)} \quad \text{(II.2)}$$

where $\alpha_M=\Phi(t_M)$ is the initial phase. The parameters A and β also implicitly depend on t_M, but the M has been dropped to emphasize that they have a constant value on the interval of analysis. For a time $\tau_m = t_m - t_M$ in the analysis interval W, the local phase is equal to $\alpha_m = \alpha_M + \beta\tau_m$.

Incorporating (II.2) in (II.1) for an even window h, leads to the identity between the Fourier phase $\arg(X(t_m,\omega_k;h))$ and the sinusoid phase α_m in the kth-channel, next to the frequency β. For a hop-size $T = t_{m+1} - t_m$, the basic phase-vocoder computes an estimate of the latter parameter as

$$\hat{\beta} = \frac{\arg(X(t_{m+1},\omega_k;h)) - \arg(X(t_m,\omega_k;h))}{T} = \frac{\Delta X}{T}. \quad \text{(II.3)}$$

The tuning of the hop-size T is a critical part of the method. The original phase vocoder described in [25] uses a one sample interval, $T = 1/F_s$. In this case, the frequency estimation requires two adjacent Fourier transforms, even if for some specific windows, such as Hann or rectangular windows, only one FT computation is made, with the second one being recursively derived from the former [32]. When the first-order phase difference is done from sample to sample, the obtained frequency estimates often display a large variance. In order to derive more efficient estimates, larger hop sizes are suggested [26]. However, this can lead to phase indetermination issues: when the phase increment βT between two successive FT is larger than 2π. The frequency estimate then becomes

$$\hat{\beta} = \frac{\Delta X + 2\pi n}{T} \quad \text{(II.4)}$$

where n is an integer, practically computed as [33]:

$$\hat{n} = \mathrm{round}\left((\omega_k T - \Delta X)/(2\pi)\right). \quad \text{(II.5)}$$

ω_k is the frequency bin next to β. A sufficient condition on T for this equation to be valid is $T < \tau_N$, or equivalently $H < N$, if H is the hop size in samples. This condition is different from the one presented in [27] and [28], where $H \leq N/4$, because here the phase vocoder is applied only to the maximal bins of the Fourier transform.

[1]Another more restricted meaning for the term phase vocoder refers only to this particular frequency estimator.

[2]in reference to the property of the phase spectrum for a symmetric analysis window

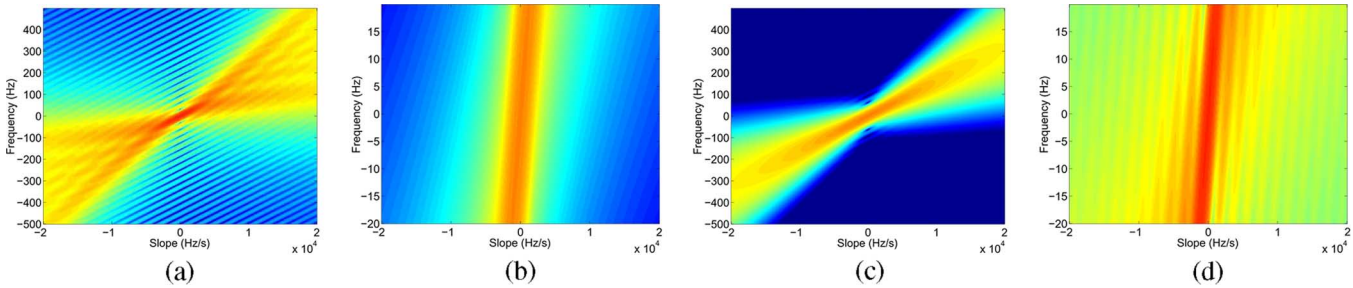

Fig. 1. Discrete quadratic phase transform of two different analysis windows. (a) Hann window; (b) Hann window (zoom); (c) rectangular window; and (d) rectangular window (zoom).

III. FM Models for the Phase Vocoder

In this section, the more elaborated quadratic phase model M02 is considered. Already widely used in frequency estimation problems (and, in particular, for frequency reassignment methods [2]), it supposes a local linear frequency (quadratic phase) and constant amplitude

$$x_{\mathrm{loc}}(\tau) \triangleq A\mathrm{e}^{j(\alpha_M+\beta_M\tau+\gamma\tau^2/2)} \tag{III.1}$$

where α_M, β_M, and γ are, respectively, the phase, frequency, and frequency change rate (FCR) for the time t_M. To emphasize that the amplitude and the FCR are constant within the analysis window, the M has been dropped. For a time t_{m_i} in the analysis interval W, the corresponding local phase is equal to $\alpha_i = \alpha_M + \beta_M\tau_{m_i} + \gamma(\tau_{m_i}^2/2)$, and the local frequency is equal to $\beta_i = \beta_M + \gamma\tau_{m_i}$.

Let us consider two frames centered in t_{m_1} and t_{m_2} such that $t_M - t_{m_1} = t_{m_2} - t_M = T/2$. From the definition of α_i, the following expression is obtained:

$$\beta_M = \frac{\alpha_2 - \alpha_1}{T}. \tag{III.2}$$

One of the differences with the previous linear model is the necessity to specify the time $\hat{t}$ at which the frequency $\hat{\beta}$ is estimated, as the frequency is not constant anymore. Here, $\hat{t} = t_M$ and $\hat{\beta}$ is an estimation of β_M.

A. Phase Error Using the FT

The other major difference with the vocoder based on the linear-phase model is that the FT becomes a biased estimator of the phase. In fact, the FCR introduces an error term Γ which depends on three parameters: the window h, the FCR γ, and the difference $\Delta\beta_i$ between the sinusoid frequency at the instant t_{m_i} and the frequency of the closest bin for the analysis window centered in t_{m_i} [34]:

$$\alpha_i = \arg\left(X\left(t_{m_i},\omega_{k_i};h\right)\right) - \arg\left(\Gamma(\Delta\beta_i,\gamma;h)\right)[2\pi] \tag{III.3}$$

$$\begin{aligned}\Gamma(\Delta\beta_i,\gamma;h) &\triangleq \sum_{n=-(N-1)/2}^{(N-1)/2} h(\tau_n)\mathrm{e}^{j\left(\Delta\beta_i\tau_n+\frac{\gamma}{2}\tau_n^2\right)} \\ &= \sum_{n=-(N-1)/2}^{(N-1)/2} h(\tau_n)\cos(\Delta\beta_i\tau_n)\mathrm{e}^{j\frac{\gamma}{2}\tau_n^2}. \end{aligned}\tag{III.4}$$

The last equality comes from a symmetry hypothesis on h. Similarly, to the linear phase model, if T is large enough, an unwrapping factor n will be needed when considering the phase difference $\alpha_2 - \alpha_1$.

It is interesting to note that the Γ function can be interpreted as the discrete quadratic phase transform (DQPT) of the window h, which is a transform used to analyze chirp signals [35], [36]. As an illustration, Fig. 1 represents the DQPT of two different analysis windows. When the window is applied to a linear chirp, the window response is translated in the frequency-FCR plane.

B. Maximum Bins Tracking

The error term $\Gamma(\Delta\beta,\gamma;h)$ is more influenced by the FCR γ than by $\Delta\beta$ when $\Delta\beta$ is less than R. This is illustrated in Fig. 1(b) and (d): for the frequency varying from $[-R,R]$ (here $R/(2\pi) \approx 15$ Hz), identical amplitude values are almost on vertical lines, i.e., the error term is dominated by the FCR influence. On the other hand, when considering larger frequency intervals, in Fig. 1(a) and (c), it can be seen that this is not true anymore; the error equally depends on both terms.

It suggests that the use of a maximum bin tracker would improve the phase vocoder. If the bins used to compute the phases α_1 and α_2 are both the closest maximum bins to the true frequency in t_{m_1} and t_{m_2}, respectively, the influence of $\Delta\beta$ will be approximately negligible [Fig. 1(b) and (d)]. Since the FCR parameter is supposed identical for both frames, the error term (III.4) in t_{m_1} and t_{m_2} will approximately cancel each other in formula (III.2).

Another obvious advantage of using maximum bins, is to reduce the influence of the noise, because the more a component is energetic, the less it will be sensitive to noise perturbation. This maximum-bin phase difference will be referred as

$$\Delta X_m \triangleq \arg\left(X\left(t_{m_2},\omega_{k_2};h\right)\right) - \arg\left(X\left(t_{m_1},\omega_{k_1};h\right)\right) \tag{III.5}$$

where ω_{k_1} (respectively, ω_{k_2}) is the frequency of the closest bin k_1 (respectively, k_2) to the sinusoid, for the time t_{m_1} (respectively, t_{m_2}).

Implementation Details: A simple tracking method can be adapted from [33]. The tracking is done locally on three consecutive frames, and the maximum frequency variation supposed in [33], corresponds here to a bound on the FCR: $\gamma \in [-\gamma_m, \gamma_m]$. Equivalently, it corresponds to a maximum bin variation $\Delta k = NT\gamma_m/4\pi F_s$, for a time interval of $T/2$. The procedure is summarized as follows.

1) Compute the zero-phased FFTs X_1, X_2, X_M for the times t_{m_1},$t_{m_2} = t_{m_1} + T$ and $t_M = t_{m_1} + T/2$.
2) Compute the maximum bin $\hat{k}_M$ for the time t_M.
3) Compute $\hat{k}_1 = \arg\max_{k\in\{k_M \pm \Delta k\}} |X(t_{m_1}, \omega_k; h)|$.
4) Compute $\hat{k}_2 = \arg\max_{k\in\{k_M \pm \Delta k\}} |X(t_{m_2}, \omega_k; h)|$.

C. Phase Corrected Vocoder

As mentioned earlier, the FT is not a direct estimator of the phase for chirp signals. A further improvement to the phase vocoder consists in correcting the Fourier phase estimation, as in [34], using the error function $\Gamma(\Delta\beta, \gamma; h)$. The estimation scheme proposed involves the following two steps:

1) estimation of the corrected phases (modulo 2π) $\hat{\alpha}_1$ and $\hat{\alpha}_2$, and the unwrapping factor $\hat{n}$;
2) estimation of β_M using the phase vocoder formula[3]

$$\hat{\beta}_M = \frac{\mathrm{mod}(\hat{\alpha}_2) - \mathrm{mod}(\hat{\alpha}_1) + 2\pi\hat{n}}{T}. \tag{III.6}$$

The function Γ requires the knowledge of the frequencies corresponding to t_{m_1} and t_{m_2} (namely β_1 and β_2). Therefore, the first step of the estimation scheme will involve a first frequency estimation for β_1 and β_2. As there is no knowledge about the FCR in this step, it is proposed to use one of the frequency estimators based on the M01 sinusoidal model. Although these estimators are biased for the FM model, it is shown in [34] that this scheme can greatly improve the precision on the phase estimates.

The parameter γ, and the unwrapping factor n can be deduced from the frequencies β_1 and β_2, using the formulas

$$\hat{\gamma} = \frac{\hat{\beta}_2 - \hat{\beta}_1}{T}$$
$$\hat{n} = \mathrm{round}\left(\frac{1}{2\pi}\left(\mathrm{mod}(\hat{\alpha}_1) - \mathrm{mod}(\hat{\alpha}_2) + \frac{\hat{\beta}_1 + \hat{\beta}_2}{2}T\right)\right)$$

In order to compute fast corrections of the STFT phase, the Γ function can be precomputed or modeled for the predefined finite intervals: $\gamma \in [0, \gamma_m]$ and $\Delta\beta \in [0, R]$. The latter interval comes from the fact that, for the FM model, the selected bin (maximum bin) corresponds to the closest bin to β. There is no need to precompute the error for the negative values of the parameters, as the function Γ is symmetric with respect to $\Delta\beta$ and antisymmetric with respect to γ.

[3]$\mathrm{mod}()$ is the modulo 2π function.

Implementation Details: The scheme of the algorithm is presented in Fig. 2. The initial frequency estimation can be done by using one-frame-based frequency estimators, such as spectrum interpolation methods. We have chosen to use the interpolation method described in [37], because of its precision and its ability to work with any window. Methods exist to estimate the FCR within a frame, but they are complex and very sensitive to noise [38], [39]. This is another motivation for using a local tracking: the greater the interval T is, the better the FCR will be estimated. The phase error function Γ has been modeled by a two-dimension lookup table. When a couple $(\Delta\beta, \gamma)$ falls between the precomputed values, $\Gamma(\Delta\beta, \gamma; h)$ is linearly interpolated from the closest elements of the table.

Algorithm: Phase corrected vocoder (PCV)

1) Maximum bin tracking: the FFTs and $\hat{k}_M$, $\hat{k}_1$ and $\hat{k}_2$ are computed (cf. Section III-B).
2) Estimation of the corrected phases:
   — Compute a first estimation $\hat{\beta}_1$ (respectively, $\hat{\beta}_2$) of the frequency in t_{m_1} (respectively, t_{m_2}).
   — Compute an estimation of the FCR:

$$\hat{\gamma} = (\hat{\beta}_2 - \hat{\beta}_1)/T.$$

   — Compute the corrected phases $\hat{\alpha}_1$, $\hat{\alpha}_2$:

$$\hat{\alpha}_i = \arg\left(X\left(t_{m_i}, \omega_{k_i}; h\right)\right) - \arg\left(\Gamma\left(\omega_{k_i} - \hat{\beta}_i, \hat{\gamma}; h\right)\right.$$

   — Compute the phase unwrapping factor:

$$\hat{n} = \mathrm{round}\left(\frac{\mathrm{mod}(\hat{\alpha}_1) - \mathrm{mod}(\hat{\alpha}_2) + .5(\hat{\beta}_1 + \hat{\beta}_2)T}{2\pi}\right).$$

3) Estimation of frequency in $\hat{t} = t_M$:

$$\hat{\beta} = \frac{\mathrm{mod}(\hat{\alpha}_2) - \mathrm{mod}(\hat{\alpha}_1) + 2\pi\hat{n}}{T}.$$

Note that this scheme could also be applied to improve the frequency precision after an initial sinusoidal tracking such as [33] and [40].

D. Reassigned Vocoder

In this section, a different approach is proposed, based on Taylor expansions of the error term in (III.3). The starting point of this approach is the phase difference formula for chirp signals [directly derived from (III.3)]:

$$T\beta_M = \Delta X_m + \arg(\Gamma_1\bar{\Gamma}_2) + 2\pi n \tag{III.7}$$

where $\Gamma_i = \Gamma(\Delta\beta_i, \gamma; h)$.

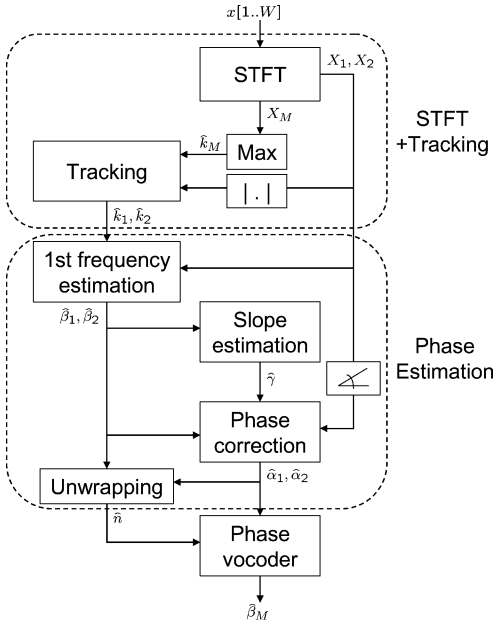

Fig. 2. Phase corrected vocoder (PCV) analysis scheme.

The first step is to express $\arg(\Gamma_1\bar{\Gamma}_2)$ as a function of β_M, the frequency to be estimated. In this process, $\Delta\beta_1$ and $\Delta\beta_2$ can be decomposed into two bounded terms B and G, described below. Let us define ω_M as the mean bin frequency and $\Delta\omega$ as half the frequency variation in bins, as follows:

$$\omega_M = \frac{\omega_{k_1} + \omega_{k_2}}{2}, \qquad \Delta\omega = \frac{\omega_{k_2} - \omega_{k_1}}{2}. \tag{III.8}$$

In addition, from the definition of the quadratic phase model (III.1), the FCR γ follows this relation:

$$\gamma\frac{T}{2} = \frac{\beta_2 - \beta_1}{2}. \tag{III.9}$$

Let $B = \beta_M - \omega_M$ and $G = \Delta\omega - \gamma(T/2)$. From the previous definitions, $\Delta\beta_i$ can be expressed as

$$\Delta\beta_1 = B + G, \quad \Delta\beta_2 = B - G \tag{III.10}$$

and since ω_{k_1} and ω_{k_2} are, respectively, the closest bins to β_1 and β_2, then $|B| < R$ and $|G| < R$, where R is the half FT precision.

The symmetry in the expression of $\Delta\beta_1$ and $\Delta\beta_2$ will simplify the remaining developments. Using first-order Taylor expansion in $G = 0$ in (III.7) leads to the following simplified expressions (see the Appendix for further details)[4]:

$$\begin{aligned}\hat{\beta} &= \frac{\Delta X_m + 2\pi n}{T} + 2\frac{\Delta\omega}{T}\Re\left(\frac{X(t_M,\omega_M;\mathcal{T}h)}{X(t_M,\omega_M;h)}\right)\\ \hat{t} &= t_M + \Re\left(\frac{X(t_M,\omega_M;\mathcal{T}h)}{X(t_M,\omega_M;h)}\right)\end{aligned} \tag{III.11}$$

[4]$\Re()$ is the real part operator.

TABLE I
MAXIMUM HOP-SIZE VALUES IN SAMPLES FOR THE RV METHOD
($N = 512$, $F_s = 16\,000$)

	Hann	Hamming	Blackman	Gaussian
$\gamma_m = 1000$	508	505	509	506
$\gamma_m = 8000$	495	492	500	494

where $\mathcal{T}h(\tau) \triangleq \tau h(\tau)$ and where $\hat{t}$ is the reassigned time. It can be observed that the frequency estimator $\hat{\beta}$ is split in a main term related to the basic vocoder and a corrective term related to the reassigned time. Note that (III.11) also indicates that the time of estimation of the basic vocoder (i.e with $\Delta\omega = 0$) is in fact the reassigned time of a frame centered on t_M for the frequency $\omega_M = \omega_{k_1}$.

The last problem to solve is the computation of the unwrapping factor n. It will be achieved using the estimator (II.5), but considering a different frequency of reference ω_M instead of ω_k [33]. As in Section II-B, this choice imposes a theoretical limit on the hop-size length of the phase vocoder, which is now discussed. From (III.7), n verifies this relation:

$$n = \frac{1}{2\pi}\left[\omega_M T - \Delta X_m + \Delta\beta_M T - \arg(\Gamma_1\bar{\Gamma}_2)\right] \tag{III.12}$$

where $\Delta\beta_M = \beta_M - \omega_M$. The chosen estimator of n is

$$\hat{n} = \text{round}\left(\frac{\omega_M T - \Delta X_m}{2\pi}\right) \tag{III.13}$$

After straightforward but tedious developments [41], it can be shown that a sufficient condition for identity between n and $\hat{n}$ is

$$H \leq N\left(1 - \frac{\Gamma_m(R,\gamma_m;h)}{\pi}\right) \tag{III.14}$$

where H is the hop-size in samples and Γ_m is the maximum value of the corrective term for the considered system parameters

$$\Gamma_m(R,\gamma_m;h) = \max_{|\Delta\beta_i|\leq R, |\gamma|\leq\gamma_m} \left|\arg(\Gamma_1\bar{\Gamma}_2)\right|. \tag{III.15}$$

When $\gamma_m = 0$, we find the classical unwrapping condition $H \leq N$. This maximum is difficult to solve analytically in the general case, but for a given set of parameters, a numerical evaluation can be done. Table I gives maximum hop-size values for various system parameters. It can be seen that the maximal theoretical hop-sizes decreases very slowly when the FCR increases. For usual applications, which use much lower hop-sizes than this limit, this means that the FCR will have no impact on the unwrapping estimation. The rectangular window cannot be used with this method (the time reassignment requires smooth functions) and is therefore not present in the table.

Implementation Details: The reassigned vocoder (RV) is implemented similarly to the PCV using three consecutive overlapping frames localized at t_{m_1}, t_M and t_{m_2} (see Fig. 3).

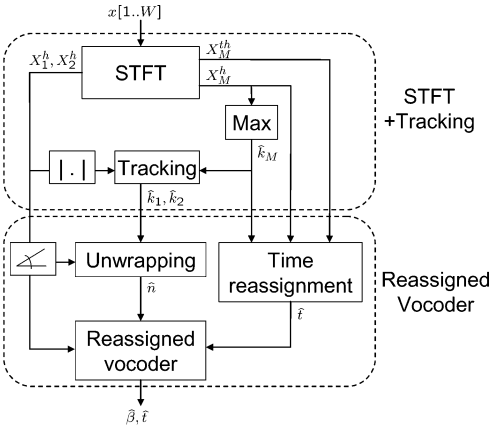

Fig. 3. Reassigned vocoder analysis scheme. X_i^h denotes the FFT computed with the window h at time t_{m_i}.

When the maximum bins k_1 and k_2 differ, the frequency $\omega_M = (k_1 + k_2)F_s/(2N)$ for the middle frame centered at t_M may not fall on a STFT bin. In order to be able to compute the reassigned time using the STFT [cf. (III.11)], k_1 or k_2 has to be moved to an adjacent bin. The new bin is chosen as the highest in amplitude between these adjacent bins.

Algorithm: Reassigned Vocoder (RV)

1) STFT computing and maximum bin tracking: Compute the FFTs using the window h and the window $\mathcal{T}h$ for the time t_M. Compute the FFTs using the window h, for the adjacent times $t_{m_1} = t_M - T/2$ and $t_{m_2} = t_M + T/2$. Compute k_M, k_1 and k_2.
2) Compute the phase unwrapping factor in two steps:

$$\hat{n}_1 = \text{round}\left(\frac{\arg(X_1) - \arg(X_M) + (\omega_{k_1} + \omega_M)T/4}{2\pi}\right)$$
$$\hat{n}_2 = \text{round}\left(\frac{\arg(X_M) - \arg(X_2) + (\omega_M + \omega_{k_2})T/4}{2\pi}\right).$$
$$\hat{n} = \hat{n}_1 + \hat{n}_2$$

3) Compute the estimated frequency:

$$\hat{\beta} = \frac{\Delta X_m + 2\pi\hat{n}}{T} + 2\frac{\Delta\omega}{T}\Re\left(\frac{X(t_M, \omega_M; \mathcal{T}h)}{X(t_M, \omega_M; h)}\right).$$

This estimation is made for the time:

$$\hat{t} = t_M + \Re\left(\frac{X(t_M, \omega_M; \mathcal{T}h)}{X(t_M, \omega_M; h)}\right).$$

In order to reduce phase unwrapping problems for the AM model case, the unwrapping factor is computed in two steps (cf. Section IV-A).

IV. AM/FM Rate Model

In this section, the more realistic AM/FM model M12 is considered. In fact, variation in frequencies are often combined with variations on amplitude in real audio signals and in particular for speech attacks and decays. The first-order AM and FM sinusoidal model can be written as

$$x_{\text{loc}}(\tau) \triangleq \mathrm{e}^{\lambda_M + \mu\tau} \cdot \mathrm{e}^{j(\alpha_M + \beta_M\tau + \gamma\tau^2/2)} \tag{IV.1}$$

where γ is the FCR, λ_M is the instantaneous log-amplitude, and μ is the log-amplitude change rate (ACR). As for the other local models, all these parameters correspond to the time t_M, and the M has been droped for γ and μ to emphasize that they are considered constant on the interval of analysis.

In this section, it will be shown that the RV is also a valid frequency estimators in the AM/FM case. The frequency reassignment and QIFFT methods are also briefly presented.

A. Application of the First-Order AM/FM to the Phase Vocoder

The PCV algorithm is not suited for the AM/FM model without using an auxiliary algorithm to estimate the ACR μ. However, the reassigned vocoder can be straightforwardly applied to this new model, as it will be shown below.

Let us define

$$\Gamma(\Delta\beta, \gamma, \mu; h) \triangleq \sum_{n=-(N-1)/2}^{(N-1)/2} h(\tau_n)\mathrm{e}^{\mu\tau_n}\mathrm{e}^{j\left(\Delta\beta\tau_n + \frac{\gamma}{2}\tau_n^2\right)}. \tag{IV.2}$$

The STFT of the signal (IV.1) can be expressed as

$$X(t, \omega; h) = \mathrm{e}^{\lambda + j\alpha}\Gamma(\Delta\beta, \gamma, \mu; h). \tag{IV.3}$$

Since the term $\arg(X(t, \omega; h))$ is independent of λ, the approach followed in Section III-D to derive (III.11) can be applied using $\Gamma(\Delta\beta, \gamma, \mu; h)$ instead of $\Gamma(\Delta\beta, \gamma; h)$.

Even if the expression (III.11) remains valid, the practical algorithm described in Section III-D presents a major drawback concerning the phase unwrapping factor estimation when the amplitude varies. According to the definition of the AM/FM rate model, the energy attributed to each frequency will be shifted depending on the ACR. The maximum of energy will no longer correspond to β_M, the sinusoid frequency for the middle of the window, as shown in Fig. 4(a). Therefore, the $\Delta\beta_i$ are no longer bounded by R, but rather by another term depending on the maximum ACR tolerated.

As a consequence, the maximum theoretical hop-sizes for the unwrapping estimation are more difficult to compute in this case. They should be lower than the values presented in Table I. Nevertheless, this problem can be minimized by using intermediate phases as it is done in the practical algorithm given in Section III-D.

B. QIFFT Algorithm

In the QIFFT method [14], [24], the first-order AM/FM model parameters are derived from the analytical formula of

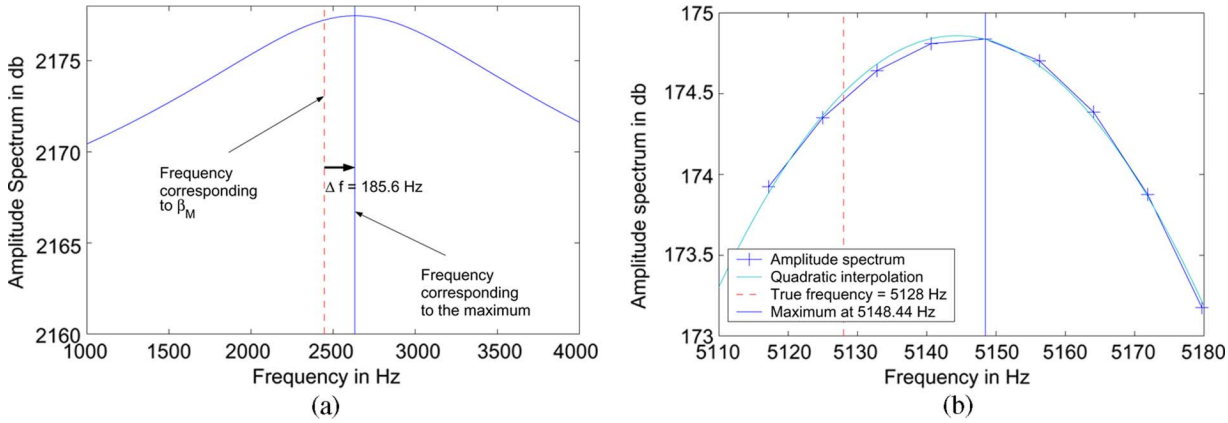

Fig. 4. Effect of high ACR and high FCR on windows. (a) Amplitude maximum is shifted for high ACR and high FCR ($\mu = 5000, \gamma = 8000$); (b) the Gaussian window's response is no longer parabolic for high modulations ($\mu = 50, \gamma = 8000$).

the Fourier Transform of a Gaussian window. It can be shown [24] that log-amplitude and phase are both quadratic functions of the Fourier frequency. Furthermore, a quadratic interpolation of both amplitude and phase will allow us to compute all the parameters of the model. Here, only the frequency estimator is studied.

Following the notations of [24], p will be the inverse of the variance of the Gaussian window: $p = 1/(2\sigma^2)$. The Gaussian window is defined by

$$w(t) = \sqrt{\frac{p}{\pi}} \mathrm{e}^{-pt^2}.$$

The log-amplitude and the phase are quadratic functions of the frequency, i.e., they are equivalent to $a_0\omega^2 + a_1\omega + a_2$ and $b_0\omega^2 + b_1\omega + b_2$, respectively. a_i and b_i are computed using parabolic interpolations on the three closest bins to the maximum of the amplitude spectrum (cf. [24] for more details). The following estimators can be derived:

$$\begin{aligned} \hat{\omega}_0 &= -\hat{a}_1/\hat{a}_0, \quad \hat{\mu} = -2p(2\hat{b}_0\hat{\omega}_0 + \hat{b}_1), \\ \hat{\gamma} &= p\hat{b}_0/\hat{a}_0, \quad \hat{\beta}_M = \hat{\omega}_0 + \hat{\mu}\cdot\hat{\gamma}/p \end{aligned}.$$

The QIFFT algorithm can be used with non-Gaussian windows, using window response adaptation. Although this adaptation has proven quite accurate, it remains only an approximation of the frequency response of the Gaussian window with the same resolution. This is why the Gaussian window has been preferred for the experiments. The tuning of the Gaussian window resolution is discussed in Section VI.

The frequency response of an infinite Gaussian window is exactly parabolic. In practice, the window is truncated and, for high ACR and FCR values, the response may not be parabolic anymore [see Fig. 4(b)]. The same problem occurs with a Hann window with the same resolution. Increasing the window length will reduce this problem.

C. Frequency Reassignment

The frequency reassignment is known to perfectly localize chirp signals [2]. A simple demonstration for the continuous Fourier transform in the FM case is presented in [7]. Using the same method, it can be shown easily that the time-frequency reassignment is also perfectly valid for the AM/FM model [41].

In keeping with the usual formulation of the reassignment, let's define $\mathcal{D}h(\tau) \stackrel{\text{def}}{=} (dh/d\tau)(\tau)$. The discrete version of the reassignment method can be defined as[5]

$$\begin{aligned} \hat{\beta} &= \omega_k - \Im\left(\frac{X(t_m, \omega_k; \mathcal{D}h)}{X(t_m, \omega_k; h)}\right) \\ \hat{t} &= t_m + \Re\left(\frac{X(t_m, \omega_k; \mathcal{T}h)}{X(t_m, \omega_k; h)}\right) \end{aligned}.$$

The discrete formulation of the reassignment introduces a small bias in the estimation which can be seen only for very high signal-to-noise ratios (SNRs) [42]. This method involves three FFT computation. It is the more complex of the method studied in this article, as the PCV and the QIFFT require one FFT, and the RV requires two FFT. The difference in complexity between all methods is mainly determined by the number and the size of the FFTs needed since they all share the same peak selection scheme.

V. Influence of Noise on the Phase Vocoder

The influence of a white noise on the basic phase vocoder is studied in [26]. A more recent reference [43] uses the same method, but presents a simpler formula applicable to any window. The results presented in [26], [43] are generalized here for the AM/FM model, defined by (IV.1).

If $S_i = S(t_{m_i}, \omega_{k_i}; h)$ and $N_i = N(t_{m_i}, \omega_{k_i}; h)$ are the Fourier transform of s and n respectively, then

$$\begin{aligned} S_i &= X_i + N_i \\ S_i &= X_i\,(1 + N_i') \end{aligned}$$

where $N_i' \triangleq N_i/X_i$. The conjugate product $S_2\bar{S}_1$ can be written as

$$S_2\bar{S}_1 = X_2\bar{X}_1(1 + Z)$$

where $Z \triangleq 1 + N_2' + \bar{N}_1' + N_2'\bar{N}_1'$.

[5] $\Im()$ is the imaginary part operator.

As it is assumed that the STFT resolves the sinusoid, from the disturbance n, it can be supposed that X_i dominates N_i in the bins close to the maximum, or equivalently that $\arg(1+Z) \approx \Im(Z) \approx \Im(N_2') - \Im(N_1')$.

$$\arg(S_2\bar{S}_1) \approx \arg(X_2\bar{X}_1) + \Im(Z).$$

Using the notations of Section IV-A, the Fourier transform of x_c for the time t_{m_1} and t_{m_2} can be put under the form

$$X_i = \mathrm{e}^{\lambda_i + j\alpha_i}\Gamma_i$$

where $\Gamma_i = \Gamma(\beta_i - \omega_{k_i}, \gamma, \mu; h)$, $i \in \{1,2\}$. From (III.2), $\alpha_2 - \alpha_1 = T\beta_M$ and $\arg(S_2\bar{S}_1)$ becomes

$$\arg(S_2\bar{S}_1) \approx T\beta_M + \arg(\Gamma_2\bar{\Gamma}_1) + 2\pi n + \Im(Z).$$

The PCV and RV methods will use estimates of $\arg(\Gamma_2\bar{\Gamma}_1)$ and n. Given that the sinusoids are well resolved by the Fourier transform, no error is done on the estimation of n, and it can be shown that the stochastic error resulting from the estimation of $\arg(\Gamma_2\bar{\Gamma}_1)$ is negligible compared to $\Im(Z)$ [41]. For both methods, the frequency estimate can be written as

$$\hat{\beta} \approx \beta + \epsilon + \frac{\Im(Z)}{T} \qquad \text{(V.1)}$$

where β is the frequency to be estimated, which is β_M for the PCV and $\beta_M + \gamma\Re(X(t_M,\omega_M;\mathcal{T}h)/X(t_M,\omega_M;h))$ for the RV estimator. $\hat{\beta}$ is the estimator for β and is given by (III.6) for the PCV and (III.11) for the RV. ϵ is the deterministic bias, which comes from the approximation described in the Appendix for the RV, and from the use of biased estimates in the first step for the PCV. For both methods, the stochastic error is approximately $\Im(Z)/T$.

The expectation of the estimators is $\beta + \epsilon$, and their variance is given by

$$\mathrm{var}(\hat{\beta}) = \frac{E\left(\Im(Z)^2\right)}{T^2} \qquad \text{(V.2)}.$$

The derivation of the variance is rather tedious and can be found in [41]

$$\mathrm{var}(\hat{\beta}) = \frac{\sinh(\mu\tau_W)}{\mu\tau_W}\frac{[\cosh(\Delta\lambda)H_0 - \cos(\Delta\Phi)H_1]}{\eta T^2|\Gamma_2\Gamma_1|} \qquad \text{(V.3)}$$

where η is the SNR, $\Delta\lambda$ and $\Delta\Phi$ are, respectively, the log-amplitude and the phase difference between X_2 and X_1. At last, H_0 and H_1 are factors depending only on the window h

$$\begin{aligned}
\eta &\triangleq \frac{\mathrm{e}^{\lambda_1+\lambda_2}}{\sigma^2}\frac{\sinh(\mu\tau_W)}{\mu\tau_W}\\
\Delta\lambda &\triangleq T\mu + \log\left(|\Gamma_2|\right) - \log\left(|\Gamma_1|\right)\\
\Delta\Phi &\triangleq T(\beta_M - \omega_M) + \arg(\Gamma_2\bar{\Gamma}_1)\\
H_0 &\triangleq \sum_{i=-N/2}^{N/2} h_i^2\\
H_1 &\triangleq \sum_{i=-(N-H)/2}^{(N-H)/2} h_{i+\frac{H}{2}}h_{i-\frac{H}{2}}\cos\left(\tau_i\left(\omega_{k_1} - \omega_{k_2}\right)\right).
\end{aligned}$$

From this equation, the variance for the AM, FM, and basic models can be deduced directly.

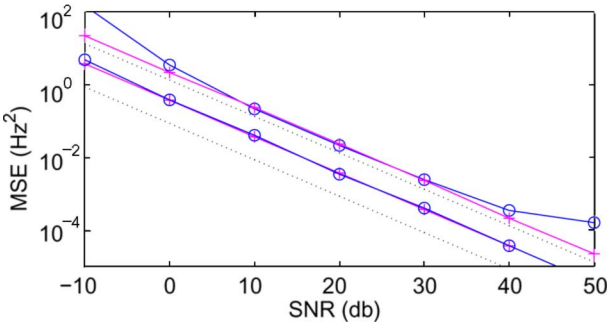

Fig. 5. Comparison of the theoretical vocoder variance ('+' markers) to the CRB (doted lines) and to the MSE of the RV method ("o" markers). Upper curves corresponds to the AM/FM model with $\mu \in [0, 100]$ and $\gamma \in [0, 8000]$, and lower curves to the FM model.

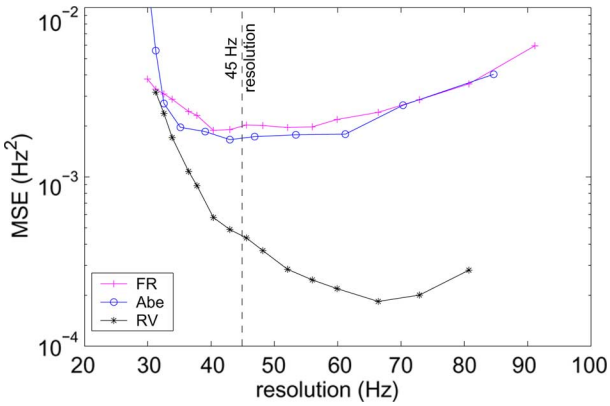

Fig. 6. Performance of the algorithms for a quadratic phase model as a function of the frequency resolution, for SNR $= 30$ db and W $= 48$ ms.

For the basic model, $\mu = 0$, $\gamma = 0$, $\Delta\lambda = 0$, and $\omega_{k_1} = \omega_{k_2} = \omega$, $\beta_1 = \beta_2 = \beta$. Equation (V.3) simplifies to

$$\mathrm{var}(\hat{\beta}) = \frac{[H_0 - \cos(\Delta\Phi)H_1]}{\eta T^2|\Gamma|^2} \qquad \text{(V.4)}$$

where

$$\eta = \frac{\mathrm{e}^{2\lambda}}{\sigma^2}, \quad H_1 = \sum_{i=-(N-H)/2}^{(N-H)/2} h_{i+\frac{H}{2}}h_{i-\frac{H}{2}},$$
$$\Delta\Phi = T(\beta_M - \omega_M), \quad \Gamma = \sum_{i=-N/2}^{N/2} h_i.$$

This last equation is the same as in [43].

Two examples are given on Fig. 5, one for the FM model (lower curves) and one for the AM/FM model (upper curves). In areas where the stochastic errors dominate, the theoretical variance match the experimental MSE of the estimators. For the AM/FM model (upper curves), biases appear at high SNRs and low SNRs. In the former case, it is caused by the deterministic error of the estimator and, in the latter case, by the tracking scheme (cf. the AM/FM case in Section VI).

VI. Evaluation

As a preamble, three important remarks can be made. The first remark concerns the peak detection step. In most exper-

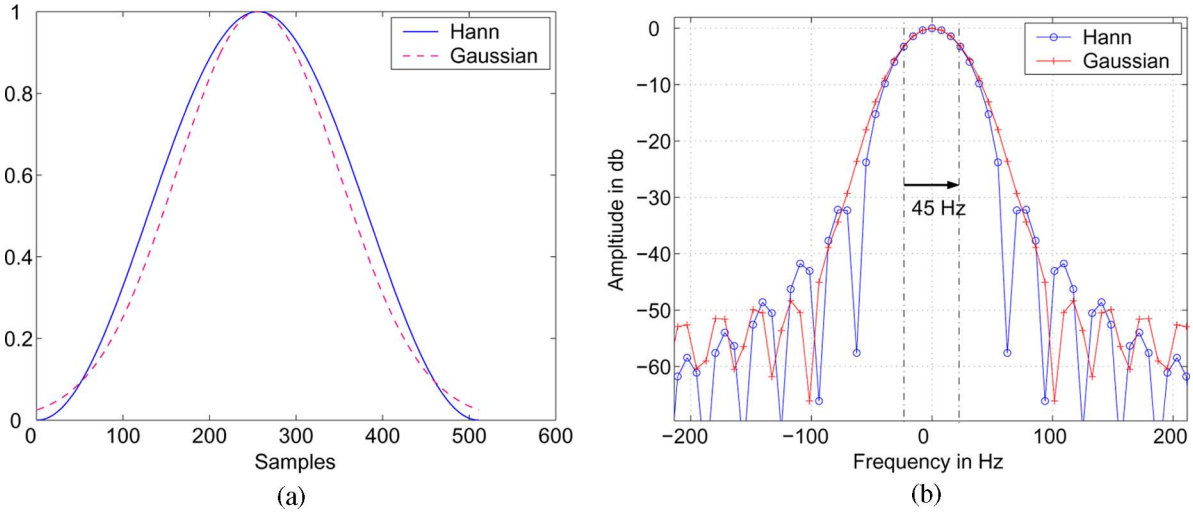

Fig. 7. Comparison of Hann and Gaussian windows for a 45 Hz resolution. (a) Hann and normalized Gaussian windows; (b) Hann and Gaussian frequency response.

imental evaluations of STFT-based frequency estimators, the highest amplitude bin will be selected as the correct peak. If this approach is very satisfactory for high SNR, it may lead to significant errors for low SNR. Since we do not aim at evaluating the bin selection algorithm, we assume, in the remaining of this paper, that the correct bins are known, i.e., the closest bins to the mean frequency of the sinusoid, β_M. To allow meaningful comparison of all methods, the correct bin is supposed to be known only for the central frame.

The second remark concerns the performance dependence of the algorithms on the STFT resolution (see Fig. 6). In fact, for a rigorous evaluation, all methods shall have the same resolution (or approximately the same). The resolution is defined as the size of the main lobe in the magnitude spectrum of the window, at -3 db from the maximum.[6] The chosen resolution, i.e., 45 Hz, shown in Fig. 6, falls in an area where all methods perform well. It corresponds to a Hann window size of 32 ms, or $N = 511$ samples for a sampling frequency of $F_s = 16\,000$, which is a classical tradeoff between time and frequency resolution for fast-varying signals such as speech. To obtain an equivalent resolution for the Gaussian window, the parameter $p = 1/(2\sigma^2)$ has to be set to $p \approx 14\,500$. The corresponding window and their frequency response are drawn for $N = 511$ in Fig. 7.

The third remark concerns the Cramér–Rao bound (CRB) [19], [44] of the frequency estimation. For a quadratic phase model, it depends on the time of estimation $\hat{t}$ of the frequency. For a white Gaussian additive noise, the CRB is minimal for the middle of the window [16] where it has the same values as the CRB of the frequency estimation for a linear phase model, i.e.,

$$\mathrm{var}(\hat{\beta}) \geq \frac{12F_s^2}{N(N^2-1)\eta} \qquad \text{(VI.1)}$$

[6]It corresponds to the root-mean-square amplitude. For a sinusoidal signal and if the maximum of amplitude is normalized to 1, $A_{\mathrm{RMS}} = 1/\sqrt{2}$. If the amplitude is measured in decibels, $20.\log_{10}(A_{\mathrm{RMS}}) \approx -3.01$ db.

where η is the SNR. For the AM/FM model, this bound becomes [17]

$$\mathrm{var}(\hat{\beta}) \geq \frac{\sinh(\mu\tau_W)}{\mu\tau_W}\frac{\epsilon_0\epsilon_4 - \epsilon_3^2}{2\eta D} \qquad \text{(VI.2)}$$

where $D = \epsilon_0\epsilon_2\epsilon_4 - \epsilon_1^2\epsilon_4 - \epsilon_1\epsilon_4^2 + 2\epsilon_1\epsilon_2\epsilon_3 - \epsilon_2^3$, $\epsilon_k = \sum_{i=-(N-1)/2}^{(N-1)/2} \tau_i^k \mathrm{e}^{2\mu\tau_i}$ and η is given by (V.3). As for the FM CRB, the bound is given for the center of the window. It is a function of μ, which is allowed to vary from one experiment to another. The CRB drawn on these experiments will be the expectation of (VI.2) given the distribution of the parameter μ. Although the CRB applies to the variance of unbiased estimators, it can be usefully compared to the mean-squared error of biased estimators.

For a fair comparison with the vocoder methods, referred as RV and PCV, the reassignment and the QIFFT will be applied using three overlapping frames (RF and QIFFT3). Maximal bins are chosen using local tracking, as described in Section III-B. The final frequency estimate is taken as the average of the estimates on the three maxima. As the QIFFT is unstable using short windows for high log-amplitude and frequency change rates as explained in Section IV-B, an estimation scheme using one long window is also presented, referred as QIFFT1.

The frequency sampling is $F_s = 16\,000$, and the resolution for each method is fixed to 45 Hz. All the estimators studied are independent of the initial phase and of the initial amplitude. The error between the true and estimated values is based on an average of 1000 experiments, using random frequencies inside [0, 8000], and random FCR and ACR. Two intervals of FCR and ACR are studied: a short interval, [0, 1000] for γ and [0, 10] for μ, and a large interval, [0, 8000] and [0, 100] respectively. An ACR of approximately 100 corresponds to an increase in amplitude by 870 dB per second, which is important but not

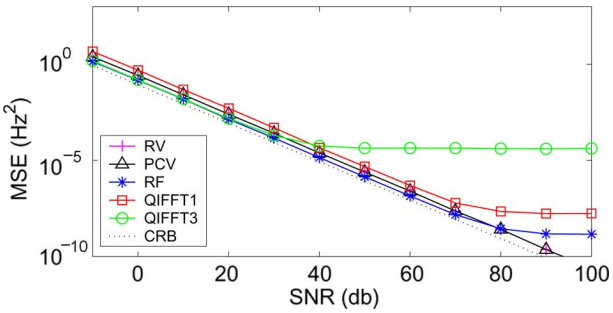

Fig. 8. Comparison of the methods for the basic model ($\gamma = 0, \mu = 0$).

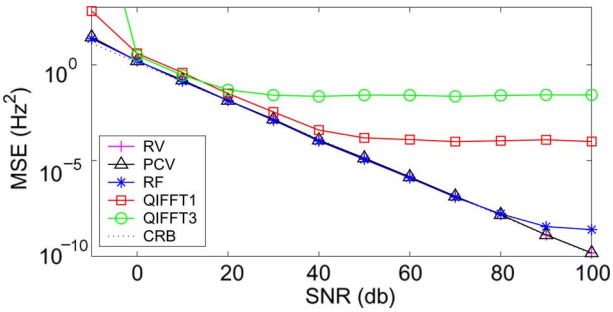

Fig. 9. Comparison of the methods for the AM model ($\mu \in [0, 100]$).

unrealistic.[7] Increasing the padding factor reduces the bias. It is especially useful for the PCV and QIFFT3 methods, which are strongly biased in the AM/FM case. When not specified, all methods are used with a padding factor of 3.

Basic and First-Order AM Models: Before studying the effect of frequency modulation, a short description of the performances for the basic model (Fig. 8) and the AM model (Fig. 9) is proposed. In both cases, it can be seen that the PCV and the RV have equal performances and are unbiased. When there is no frequency variation the estimators are identical, and equal to the basic phase vocoder frequency estimator. In the basic case, all estimators are close to the CRB. The remaining difference comes from the use of nonrectangular windows. In the short region of interest (i.e., SNR $\in [-10, 40]$), the QIFFT3 performs slightly better than the others. The bias of the QIFFT methods comes from the truncation of the Gaussian window (cf. Section IV-B). When the window is larger (method QIFFT1), the bias is reduced. For the reassignment window, the bias comes from the use of discrete FT instead of continuous FT and appears at 80 dB [42].

In the AM case, higher MSE is obtained for all estimators. However, the reassignment and the phase vocoder performances are now almost equal to the CRB. As mentioned in Section IV-B, the frequency response of the Gaussian window is no longer parabolic for high ACR, leading to important bias, seen for high SNRs. For low SNR, the QIFFT algorithm becomes unstable. This error is due to the propagation of the high-order parameters errors (here the ACR) to the lower order parameters (here the frequency), which is a known issue for this kind of estimators [16], [35]. The same problem will appear for the FM model, and for the AM/FM model.

First-Order FM Model: This section will compare the estimator presented for the first-order FM model. The experiments are done for $W = 767$ and for a FCR inside [0, 8000] (Fig. 10).

Fig. 10(a) shows the successive improvements that can be done on the long-term phase vocoder (LV). In this particular experiments, the zero-padding factor is 1 for all the methods. The upper curve corresponds to the standard long vocoder and illustrates the significant bias for high FCR signals. A first improvement is obtained when the reassigned time is used for the time of estimation of the LV. It corresponds to a first-order approximation with $\Delta\omega = 0$, as shown in Section III-D, (III.11). Using the closest maximum bins, as described in Section III-B, leads to a greater improvement. Better performances are obtained with the PCV estimation algorithm, and this especially for high SNRs. The remaining bias is mainly caused by the bias on the FCR estimates. At last the RV, combining the use of maximum bins and time reassignment, strongly improves performances for high SNRs. The estimator is in fact almost unbiased for the considered intervals of FCR.

Fig. 10(b) is a comparison of the two proposed schemes with the state-of-the-art estimators described in Section IV. The frequency reassignment and the QIFFT perform equally well for SNRs above 0 dB. For the QIFFT, and for very low SNRs, the phenomenon of error propagation from the high-order parameters (here the FCR) appears again. For the short region of interest SNR $\in [0, 40]$, all methods perform equally with a slight advantage to the QIFFT.

First-Order AM/FM Model: In this section, the methods are compared for the first-order AM/FM model. In Fig. 11(d), the FCR is inside [0, 1000] and the ACR inside [0, 10], which corresponds to sinusoids varying moderately. For all other figures, FCR is inside [0, 8000] and ACR inside [0, 100].

Fig. 11(a) and (c) illustrates the performances obtained for different window analysis sizes. For a small hop-size [Fig. 11(c)], the reassigned vocoder uses identical bins ($k_1 = k_2 = k_M$), because the frequency variation is very small in this case. Its bias almost disappears and its performances are almost identical compared to the frequency reassignment: for both methods the bias appears around 80 dB.

For large hop-sizes and low SNRs, the three-frame estimation scheme becomes slightly unstable. Actually, the difficulty of the task increases when considering larger hop-sizes, because the amplitude modulation shifts the distribution of the sinusoid energy to one of the edges of the window W. The other edge has a lower energy, overwhelmed by noise in the low SNRs cases, causing difficulties to the tracking scheme. This effect can be slightly seen in Fig. 11(a) for the RV and RF curves. A 16 ms hop-size is a good compromise, keeping good performances with a reasonable length between frames. The bias of the reassignment is identical as in previous experiments, whereas the bias of the RV method slightly increases. In the short region of interest SNR $\in [0, 40]$, the RV and the RF methods are both close to the CRB with a slight advantage to RV method (Fig. 6).

[7]For example, a sharp trumpet attack can exhibit an amplitude increase by 30 dB in less than 30 ms, leading to an increase by about 1000 dB per second. In speech, during fast transition between phonemes, the pitch can rise at an FCR of 5000 Hz/s, not to mention the harmonics.

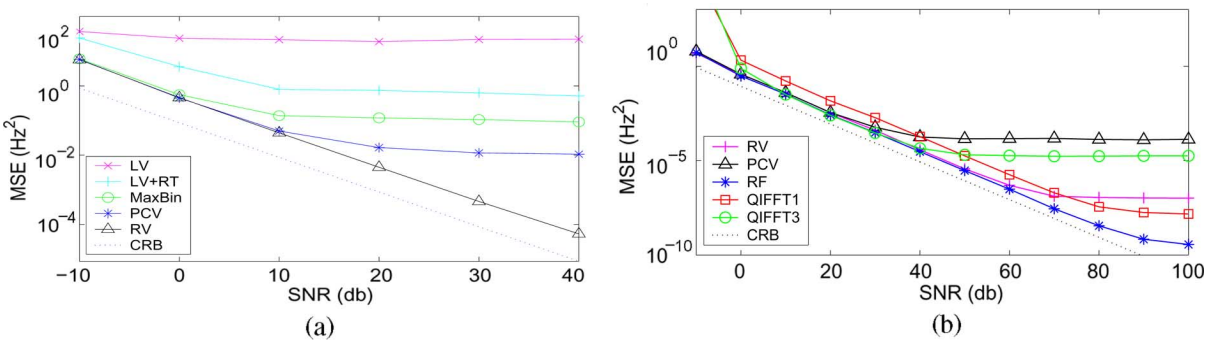

Fig. 10. Comparison of the different methods for the FM model ($\gamma \in [0, 8000]$). (a) Comparison of vocoder-based methods: Long vocoder (LV); LV using maximum bin (MaxBin); LV with reassigned time (LV + RT); and RV and PCV. (b) Comparison of the PCV, the reassigned frequency (RF), the QIFFT method, and the RV.

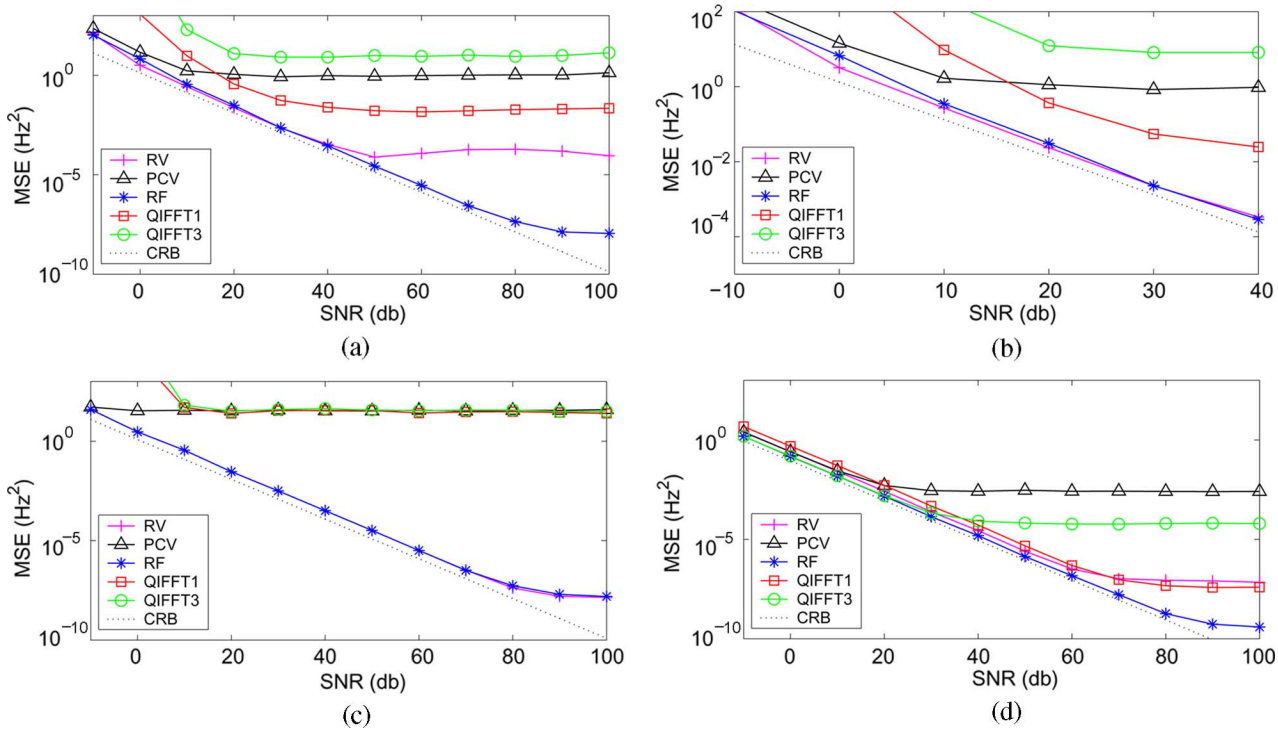

Fig. 11. Performance of the three methods for an AM/FM model. (a) $\gamma \in (0, 8000), \mu \in (0, 100), W = 767, T = 16$ ms; (b) $\gamma \in (0, 8000), \mu \in (0, 100), W = 767, T = 16$ ms (short region); (c) $\gamma \in (0, 8000), \mu \in (0, 100), W = 513, T = 0.1$ ms; and (d) $\gamma \in (0, 1000), \mu \in (0, 10), W = 767, T = 16$ ms.

As mentioned in Section VI, too small a Gaussian window does not have a parabolic response, leading to parameter estimation problems. When increasing the window size, the QIFFT approach is more accurate but remains unstable for low SNRs and is outperformed by the other two estimators. In this case, the PCV estimation is no longer valid, because it does not take into account amplitude variations III-C. Fig. 11(d) shows that all methods perform well on slowly varying sinusoids, with a slight advantage to the QIFFT algorithm (QIFFT3), in the short region of interest.

VII. Conclusion

This paper presented a study of the phase vocoder in the case of two common models, the chirp (or FM) model and the first-order AM/FM model. Two methods of estimation using the phase vocoder framework were derived, the phase-corrected vocoder, working for frequency modulation only, and the reassigned vocoder, working for both models. Performance analysis showed that the reassigned vocoder is comparable to the reassignment, but with a slight decrease in complexity, as one less STFT is needed.

Several extensions of this work can be considered. First, it should be necessary to consider more realistic signals involving multiple sinusoids. Second, further work is needed concerning the estimation of the frequency and log-amplitude change rates to extend the low complexity PCV approach to the more general case of AM/FM signals. Finally, in a similar way as for the phase vocoder, the influence of frequency varying models on other well-known frequency estimators, such as the Discrete fourier spectrum interpolators using phase, should be explored.

APPENDIX

The first-order Taylor expansion in $G = 0$ of Γ_1 and Γ_2 is given by

$$\begin{aligned}\Gamma_1 &= \Gamma(B,\gamma;h) + G\Gamma'(B,\gamma;h) + \epsilon_1 \\ \Gamma_2 &= \Gamma(B,\gamma;h) - G\Gamma'(B,\gamma;h) + \epsilon_2\end{aligned}$$

where ϵ_1 and ϵ_2 are the Lagrange remainders. The frequency derivation property of the STFT leads to

$$\begin{aligned}\Gamma_1 &= \Gamma(B,\gamma;h) + jG\Gamma(B,\gamma;\mathcal{T}h) + \epsilon_1 \\ \Gamma_2 &= \Gamma(B,\gamma;h) - jG\Gamma(B,\gamma;\mathcal{T}h) + \epsilon_2.\end{aligned}$$

For an order 1 Taylor expansion in 0 of the argument function, we obtain

$$\arg(\Gamma_1\bar{\Gamma}_2) = 2G\Re\left(\frac{\Gamma(B,\gamma;\mathcal{T}h)}{\Gamma(B,\gamma;h)}\right) + \epsilon.$$

This approximation has proven to be quite accurate for the intervals of parameter considered. Indeed the deterministic bias in the experimental section is seen only below 40 dB and with a magnitude of 10^{-2} Hz in average for $\mu \in [0,100]$ and $\gamma \in [0,8000]$ [cf. Fig. 11(a)]. See [41] for more details.

$\Re(\Gamma(B,\gamma;\mathcal{T}h)/\Gamma(B,\gamma;h))$ is in fact equivalent to the discrete version of the reassigned time. Indeed, the STFT can be rewritten as a function of Γ

$$\begin{aligned}X(t_M,\omega_M;\mathcal{T}h) &= Ae^{j\alpha_M}\Gamma(B,\gamma;\mathcal{T}h) \\ X(t_M,\omega_M;h) &= Ae^{j\alpha_M}\Gamma(B,\gamma;h)\end{aligned}$$

where $\mathcal{T}h(\tau) \triangleq \tau.h(\tau)$. We can therefore conclude that

$$\Re\left(\frac{\Gamma(B,\gamma;\mathcal{T}h)}{\Gamma(B,\gamma;h)}\right) = \Re\left(\frac{X(t_M,\omega_M;\mathcal{T}h)}{X(t_M,\omega_M;h)}\right)$$

and

$$\arg(\Gamma_1\bar{\Gamma}_2) = 2G\Re\left(\frac{X(t_M,\omega_M;\mathcal{T}h)}{X(t_M,\omega_M;h)}\right) + \epsilon.$$

Using the previous expression in (III.7)

$$T\beta_M = \Delta X_m + 2\pi n + 2G\Re\left(\frac{X(t_M,\omega_M;\mathcal{T}h)}{X(t_M,\omega_M;h)}\right) + \epsilon.$$

Replacing G by its definition leads to

$$\begin{aligned}\beta_M + \gamma\Re\left(\frac{X(t_M,\omega_M;\mathcal{T}h)}{X(t_M,\omega_M;h)}\right) &= \frac{\Delta X_m + 2\pi n}{T} \\ &\quad + 2\frac{\Delta\omega}{T}\Re\left(\frac{X(t_M,\omega_M;\mathcal{T}h)}{X(t_M,\omega_M;h)}\right) + \frac{\epsilon}{T}.\end{aligned}$$

The left part of this expression is the frequency for the time: $\hat{t} = t_M + \Re(X(t_M,\omega_M;\mathcal{T}h)/X(t_M,\omega_M;h))$. The right part is the vocoder estimator corrected by a term depending on the reassigned time and on $2\Delta\omega/T$, which can be interpreted as a first FCR estimate using frequency bins.

REFERENCES

[1] D. C. Rife and R. R. Boorstyn, "Multiple-tone parameter estimation from discrete-time observations," *Bell Syst. Tech. J.*, vol. 55, no. 9, pp. 1389–1410, 1976.

[2] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representation by the reassignment method," *IEEE Trans. Signal Process.*, vol. 43, no. 5, pp. 1068–1088, May 1995.

[3] M. R. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 3, pp. 364–374, Jun. 1981.

[4] M. Abe and J. O. Smith, III, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," presented at the Audio Engineering Soc. 117th Convention, San Francisco, CA, Oct. 28–31, 2004.

[5] B. G. Quinn, "Estimating frequency by interpolation using Fourier coefficients," *IEEE Trans. Signal Process.*, vol. 42, no. 5, pp. 1264–1268, May 1994.

[6] M. Macleod, "Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 141–148, Jan. 1998.

[7] M. Betser, P. Collen, B. David, and G. Richard, "Review and discussion on STFT-based frequency estimation methods," presented at the Audio Engineering Soc. 120th Convention, Paris, France, May 20–23, 2006.

[8] P. Stoica and A. Nehorai, "Statistical analysis of two non-linear least-squares estimators of sine wave parameters in the colored noise case," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1988, pp. 2408–2411.

[9] A. Choi, "Real-time fundamental frequency estimation by least-square fitting," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 2, pp. 201–205, Mar. 1997.

[10] R. Roy *et al.*, "Esprit: A subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1340–1342, Oct. 1986.

[11] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[12] K. Hermus, W. Verhelst, and P. Wambacq, "Psychoacoustic modeling of audio with exponentially damped sinusoids," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1821–1824.

[13] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech Audio*, vol. 12, no. 2, pp. 121–132, Mar. 2004.

[14] G. Peeters and X. Rodet, "Sinola: A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum," presented at the Int. Computer Music Conf. (ICMC), Beijing, China, Oct. 22–28, 1999.

[15] S. Peleg and B. Porat, "Linear FM signal parameter estimation from discrete-time observations," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 27, pp. 607–614, Jul. 1991.

[16] P. M. Djuric and S. M. Kay, "Parameter estimation of chirp signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 12, pp. 2118–2126, Dec. 1990.

[17] G. Zhou, G. Giannakis, and A. Swami, "On polynomial phase signals with time-varying amplitudes," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 848–861, Apr. 1996.

[18] R. Badeau, R. Boyer, and B. David, "Eds parametric modeling and tracking of audio signals," in *Proc. Digital Audio Effects (DAFx)*, Hamburg, Germany, Sep. 26–28, 2002, pp. 139–144.

[19] D. C. Rife and R. R. Boorstyn, "Single-tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, Sep. 1974.

[20] J. Wolcin, "Maximum *a posteriori* estimation of narrowband signal parameters," *J. Acoust. Soc. Amer.*, vol. 68, no. 1, pp. 174–178, Jul. 1980.

[21] S. Saha and S. M. Kay, "Maximum likelihood parameter estimation of superimposed chirps using Monte Carlo importance sampling," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 224–230, Feb. 2002.
[22] T. Abotzoglou, "Fast maximum likelihood joint estimation of frequency and frequency rate," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1986, pp. 1409–1412.
[23] B. Friedlander and J. Francos, "Estimation of amplitude and phase of non-stationary signals," in *Proc. 27th Asilomar Conf. Signals, Syst., Comput.*, Apr. 1993, pp. 848–861.
[24] M. Abe and J. O. Smith, III, "AM/FM rate estimation for time-varying sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Philadelphia, PA, Mar. 18–23, 2005, vol. 3, pp. 201–204.
[25] J. Flanagan and R. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, pp. 1493–1509, Nov. 1966.
[26] M. S. Puckette and J. C. Brown, "Accuracy of frequency estimate using the phase vocoder," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 166–176, Mar. 1998.
[27] J. Laroche and M. Dolson, "New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects," in *Proc. IEEE Workshop Applications Signal Process. Audio Acoust.*, Oct. 1999, pp. 91–94.
[28] M. S. Puckette, "Phase-locked vocoder," in *Proc. IEEE ASSP Workshop Applications Signal Process. Audio Acoust.*, 1995.
[29] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, pp. 174–215, 1995.
[30] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 236–243, Apr. 1984.
[31] J. M. Grey and J. A. Moorer, "Perceptual evaluation of synthesized musical instrument tone," *J. Acoust. Soc. Amer.*, vol. 62, pp. 454–462, Aug. 1977.
[32] J. Brown and M. Puckette, "A high resolution fundamental frequency determination based on phase changes of the Fourier transform," *J. Acoust. Soc. Amer.*, vol. 94, pp. 662–667, Aug. 1993.
[33] R. McAulay and T. 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.
[34] M. Betser, P. Collen, and J.-B. Rault, "Accurate FFT-based phase estimation for chirp-like signals," presented at the Audio Eng. Soc. 120th Conv., Paris, France, May 20–23, 2006.
[35] M. Z. Ikram, K. Abed-Meraim, and Y. Hua, "Fast discrete quadratic phase transform for estimating the parameters of chirp signals," in *Proc. IEEE Conf. Signals, Syst., Comput.*, Nov. 1996, vol. 1, pp. 798–802.
[36] X. Xia, "Discrete chirp-Fourier transform and its application to chirp rate estimation," *IEEE Trans. Signal Process.*, vol. 48, no. 11, pp. 3122–3133, Nov. 2000.
[37] M. Betser, P. Collen, and G. Richard, "Frequency estimation based on adjacent DFT bins," presented at the EUSIPCO, Florence, Italy, Sep. 4–8, 2006.
[38] A. Röbel, "Estimating partial frequency and frequency slope using reassignment operators," in *Proc. Int. Computer Music Conf.*, Sep. 2002, pp. 122–125.
[39] A. S. Master and Y.-W. Liu, "Non-stationary sinusoidal modeling with efficient estimation of linear frequency chirp parameters," presented at the IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Hong Kong, China, Apr. 6–10, 2003.
[40] X. Serra, *Musical Sound Modeling With Sinusoids Plus Noise*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. Lisse, Holland: Swets & Zeitlinger, 1997.
[41] M. Betser, P. Collen, B. David, and G. Richard, "Experimental and theoretical complements to the article 'Estimation of frequency for AM/FM models using the phase vocoder framework,'" GET Tech. Rep., 2007.
[42] S. Hainsworth and M. Macleod, "Time-frequency reassignment: Measures and uses," in *Proc. Cambridge Music Process. Colloq.*, 2003, p. 36.
[43] S. Abeysekera and K. Padhi, "An investigation of window effects on the frequency estimation using the phase vocoder," *IEEE Trans. Audio Speech Signal Process.*, vol. 14, no. 4, pp. 1432–1439, Jul. 2006.
[44] S. M. Kay, *Estimation Theory*, A. V. Oppenheim, Ed. Upper Saddle River, NJ: Prentice-Hall, 1993, 07458.

Michael Betser received the Master's degree in computer engineering from the EISTI, (Ecole Internationale des Sciences du Traitement de l'Information), Cergy-Pontoise, France, in 2002. He is currently working towards the Ph.D. degree at the TECH/IRIS (Image, Rich media, new Interactions and hyperlanguageS) laboratory of France Telecom R&D, Rennes, France.

From 2002 to 2004, he worked as an Audio Signal Processing Engineer at the IRISA (Institut de Recherche en Informatique et en Système Automatiques) in the METISS team. His research interests are in digital signal processing and audio indexing systems.

Patrice Collen received the Ph.D. degree in signal processing from the ENST of Paris, France, in 2002.

Since 2003, he has been a Research Engineer for France Telecom, Rennes, France. His research interests include audio analysis, audio indexing, and signal processing.

Gaël Richard (M'02–SM'06) received the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1990, the Ph.D. degree in speech synthesis from LIMSI-CNRS, University of Paris-XI, in 1994, and the Habilitation à Diriger des Recherches degree from the University of Paris XI in 2001.

After the Ph.D., he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the speech processing group of Prof. J. Flanagan, where he explored innovative approaches for speech production. Between 1997 and 2001, he successively worked for Matra Nortel Communications, Bois d'Arcy, France, and for Philips Consumer Communications, Montrouge, France. In particular, he was the Project Manager of several large-scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, GET-Télécom Paris (ENST), where he is now full Professor in audio signal processing and Head of the Audio, Acoustics and Waves research group. He is coauthor of over 70 papers and inventor in a number of patents. He is also an expert for the European Commission in the field of audio signal processing and man–machine interfaces.

Prof. Richard is a member of EURASIP and Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.

Bertrand David (M'06) was born on March 12, 1967, in Paris, France. He received the M.Sc. degree from the University of Paris-Sud, France, in 1991; the Agrégation, a competitive French examination for the recruitment of teachers, in the field of applied physics, from the Ecole Normale Supérieure (ENS), Cachan, France; and the Ph.D. degree from the University of Paris 6, France, in 1999, in the fields of musical acoustics and signal processing of musical signals.

He formerly taught in a graduate school in electrical engineering, computer science, and communication. He also carried out industrial projects aimed at embarking a low-complexity sound synthesizer. Since September 2001, he has worked as an Associate Professor with the Signal and Image Processing Department, GET-Télécom Paris (ENST), France. His research interests include parametric methods for the analysis/synthesis of musical and mechanical signals, music information retrieval, and musical acoustics.