ANECHOIC PHASE ESTIMATION FROM REVERBERANT SIGNALS

A. Belhomme, Y. Grenier, R. Badeau

LTCI, CNRS, Télécom ParisTech,
Université Paris-Saclay,
75013 Paris, France

E. Humbert

Invoxia,
Audio Lab,
92130 Issy-les-Moulineaux, France

ABSTRACT

Most dereverberation methods aim to reconstruct the anechoic magnitude spectrogram, given a reverberant signal. Regardless of the method, the dereverberated signal is systematically synthesized with the reverberant phase. This corrupted phase reintroduces reverberation and distortion in the signal. This is why we intend to also reconstruct the anechoic phase, given a reverberant signal. Before processing speech signals, we propose in this paper a method for estimating the anechoic phase of reverberant chirp signals. Our method presents an accurate estimation of the instantaneous phase and improves objective measures of dereverberation.

***Index Terms*—** Dereverberation, phase, reassignment, sinusoidal modeling.

1. INTRODUCTION

When a sound is emitted in an enclosed space, the microphone does not only capture the output of the source: all the paths the sound may follow, from the source to the microphone, are added to the direct one and produce reverberation. Whereas a soft reverberation may be desired to color a sound or to give a feeling of space [1], strong reverberation is unpleasant. Indeed, it damages speech intelligibility and quality when human beings are concerned and it reduces automatic speech recognition performance of machines [2].

If the room impulse response (RIR) is known, one can invert it to cancel the reverberation, thus there is no problem of phase [3, 4]. However, these cancellation methods are highly dependent on the speaker position and require a high computing power. Hence, one prefers to use suppression methods, which aim to estimate the magnitude spectrogram of the late reverberation and remove it from the magnitude spectrogram of the reverberant signal. To do so, existing methods are based on a linear prediction model [5], on a stochastic model of the RIR [6, 2], or more recently on deep neural networks [7].

However, these suppression methods present a main drawback: once the dereverberated magnitude spectrogram is computed, the reverberant phase is used to synthesize the dereverberated signal. We observed that using this corrupted phase reintroduces reverberation and distortion in the signal; this fact was also highlighted in [7]. The idea of modeling the phase has recently been proposed in source separation literature [8], where a similar problem occurs: the phase of the mixture is used to synthesize the source signals.

For speech enhancement, *Deleforge et al.* proposed an adaptation of the K-SVD dictionary method to consider the phase in a noise suppression task [9]. But for dereverberation, we were only able to find a post-processing step in [10] which constrains a harmonic structure in the dereverberated spectrogram. This phase modification enables a slight improvement of the Perceptual Evaluation of Speech Quality (PESQ, [11]) and motivates a proper estimation of the anechoic phase.

In this paper, we propose a method for estimating the phase of an anechoic signal given its reverberant version. Before working on speech we started to study chirp signals, which will be used to model voiced signals. Section 2 derives the effect of reverberation on the Hilbert phase and presents a first estimator. In Section 3, we adapt this estimator to a Short Time Fourier Transform (STFT) framework to be able to deal with multicomponent signals. The evaluation method and results are detailed in Section 4. Finally, in Section 5 some conclusions are drawn and outlooks are proposed for future work.

2. INFLUENCE OF REVERBERATION ON THE HILBERT PHASE

As we study chirp signals in the first place, we derive the expression of the Hilbert phase of a reverberant chirp, in order to highlight the influence of the reverberation. The influence on the phase is also motivated by [12], where the authors added the Hilbert phase of reverberant signals as a feature in a classification and regression tree (CART, [13]) to estimate their level of reverberation. They found that phase was one of the top 10 most important features, among 300, for the CART algorithm.

2.1. Models and notations

We consider an anechoic chirp signal $s(t)$ of T seconds

$$s(t) = \cos\left(\varphi(t)\right),\ t \in [0, T] \tag{1}$$

and the instantaneous phase $\varphi(t)$

$$\varphi(t) = 2\pi(f_d t + \frac{\dot{f}}{2}t^2) + \varphi_0,\ t \in [0, T] \tag{2}$$

with φ_0 chosen between 0 and 2π. This phase results in an instantaneous frequency $f(t)$ with a linear variation from the starting frequency f_d Hz with a rate of $\dot{f}$ Hz/s:

$$f(t) = \frac{1}{2\pi}\frac{d\varphi}{dt}(t) = f_d + \dot{f}t,\ t \in [0, T]. \tag{3}$$

The RIR $h(t)$ of length T_h is modeled with the stochastic model introduced in [14]:

$$h(t) = b(t)p(t) \tag{4}$$

with $b(t) \sim \mathcal{N}(0, \sigma^2)$ a centered Gaussian white noise of variance σ^2 and $p(t) = e^{-\delta t}\mathbb{1}_{[0,T_h]}$, where $\mathbb{1}_{[0,T_h]}$ denotes the indicator function of the interval $[0, T_h]$.

The room parameter δ is directly linked to the reverberation time RT_{60}, which is the required time to observe a 60 dB decay of the reverberant energy [15], through $\delta = \frac{3\log(10)}{\mathrm{RT}_{60}}$. Thus, we model the reverberant chirp $y(t)$ as the convolution of $s(t)$ and $h(t)$:

$$y(t) = (h * s)(t). \tag{5}$$

We denote $\widehat{y}(t)$ the Hilbert transform of a real signal $y(t)$, corresponding to the Cauchy principal value (p.v.) of

$$\widehat{y}(t) = \text{p.v.}\int_{-\infty}^{\infty}\frac{y(\tau)}{\pi(t-\tau)}d\tau.$$

Then, the Hilbert phase of the reverberant signal is obtained with

$$\varphi_{\text{rev}}(t) = \arctan\left(\frac{\widehat{y}(t)}{y(t)}\right). \tag{6}$$

2.2. Estimator with the Hilbert transform

We are interested in the instantaneous frequency of the reverberant signal:

$$f_{\text{rev}}(t) = \frac{1}{2\pi}\frac{d\varphi_{\text{rev}}}{dt}(t) = \frac{1}{2\pi}\frac{\frac{d\widehat{y}}{dt}(t)y(t) - \widehat{y}(t)\frac{dy}{dt}(t)}{\widehat{y}(t)^2 + y(t)^2}. \tag{7}$$

Firstly, we approximate the mathematical expectation of the ratio in (7) as the ratio of the mathematical expectations[1]. Thus, we define $\overline{f_{\text{rev}}(t)}$ as:

$$\overline{f_{\text{rev}}(t)} = \frac{1}{2\pi}\frac{\mathbb{E}\left[\frac{d\widehat{y}}{dt}(t)y(t) - \widehat{y}(t)\frac{dy}{dt}(t)\right]}{\mathbb{E}\left[\widehat{y}(t)^2 + y(t)^2\right]}. \tag{8}$$

The Hilbert transform behaves nicely with the convolution operator as $\widehat{(h*s)}(t) = (h * \widehat{s})(t) = (\widehat{h} * s)(t)$. We can then easily obtain the following equations:

$$\widehat{y}(t) = \int_\tau \widehat{s}(t-\tau)h(\tau)d\tau,$$

$$\frac{d\widehat{y}}{dt}(t) = \int_\tau \frac{d\widehat{s}}{dt}(t-\tau)h(\tau)d\tau.$$

Let us detail the first term in (8):

$$\mathbb{E}\left[\frac{d\widehat{y}}{dt}(t)y(t)\right] =$$

$$\mathbb{E}\left[\int_{\tau_1}\int_{\tau_2}\frac{d\widehat{s}}{dt}(t-\tau_1)s(t-\tau_2)p(\tau_1)p(\tau_2)b(\tau_1)b(\tau_2)d\tau_1 d\tau_2\right]$$

$$= \sigma^2\left(\left(\frac{d\widehat{s}}{dt}s\right) * p^2\right)(t)$$

since $\mathbb{E}[b(\tau_1)b(\tau_2)] = \sigma^2\delta_{\tau_1\tau_2}$ with δ_{ij} the *Kronecker* delta. Likewise, we have:

$$\mathbb{E}\left[\widehat{y}(t)\frac{dy}{dt}(t)\right] = \sigma^2\left(\left(\widehat{s}\frac{ds}{dt}\right) * p^2\right)(t),$$

$$\mathbb{E}\left[y^2(t)\right] = \sigma^2(s^2 * p^2)(t),$$

$$\mathbb{E}\left[\widehat{y}^2(t)\right] = \sigma^2(\widehat{s}^2 * p^2)(t).$$

Hence we can rewrite $\overline{f_{\text{rev}}(t)}$ as:

$$\overline{f_{\text{rev}}(t)} = \frac{1}{2\pi}\frac{\left(p^2 * z_1\right)(t)}{\left(p^2 * z_2\right)(t)} \tag{9}$$

with

$$z_1(t) = \frac{d\widehat{s}}{dt}(t)s(t) - \widehat{s}(t)\frac{ds}{dt}(t), \tag{10}$$

$$z_2(t) = s^2(t) + \widehat{s}^2(t). \tag{11}$$

For a cosine with fixed frequency f_0, i.e. $s(t) = \cos(2\pi f_0 t)$, we get $\hat{s}(t) = \sin(2\pi f_0 t)$ and then $\overline{f_{\text{rev}}(t)} = f_0$. There is no influence of the reverberation, as well as it is hard to notice reverberation when hearing stationary signals.

For the chirp signal defined in Section 2.1, we assume that $\hat{s}(t) = \sin(\varphi(t))$ still approximately holds, which has been confirmed experimentally on a wide range of $\dot{f}$ (for the exact theoretical conditions the reader can refer to the *Bedrosian's Theorem* in [16]). Thus, after straightforward calculations, equations (9) to (11) yield:

$$\overline{f_{\text{rev}}(t)} = \underbrace{f(t)}_{\text{signal}} - \underbrace{\dot{f}\left(\frac{1}{2\delta} + \frac{\min(t, T_h)}{1 - e^{2\delta\min(t,T_h)}}\right)}_{\text{room influence}}. \tag{12}$$

[1] This assumption has been confirmed experimentally on a wide range of $\dot{f}$ and RT_{60}, with RIR simulated with (4), resulting in less than 0.5 % relative error on the frequency.

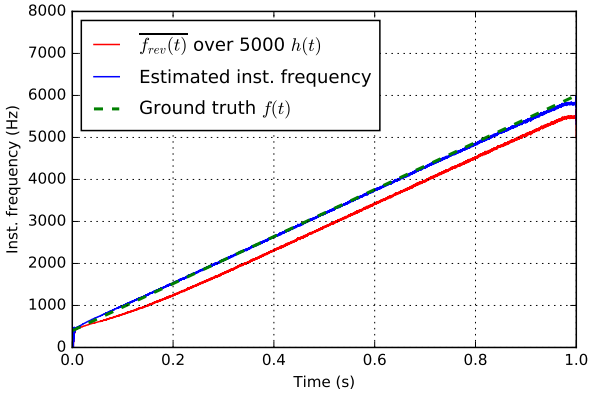

Fig. 1. Instantaneous frequency estimation

On a toy example with $T = 1.0$ s and $\text{RT}_{60} = 0.8$ s, we have plotted $f(t)$, $\overline{f_{\text{rev}}(t)}$ and $\overline{f_{\text{rev}}(t)}+\dot{f}\left(\frac{1}{2\delta}+\frac{\min(t,T_h)}{1-e^{2\delta\min(t,T_h)}}\right)$, estimating $\overline{f_{\text{rev}}(t)}$ by computing empirical means on 5000 realizations of $h(t)$ with the same RT_{60}. We can see in Figure 1 that the instantaneous frequency is well estimated.

However we do not have access to $\overline{f_{\text{rev}}(t)}$ in practice, so we approximate it with a temporal smoothing. To do so, we experimentally verified that $\overline{f_{\text{rev}}(t)}$ can be accurately estimated by means of a *Savitzky-Golay* filter h_{SG} [17] of order 1 and size N_{SG}.

Thus, given a reverberant chirp signal we can estimate the instantaneous frequency of the anechoic signal with

$$\tilde{f}(t) = (h_{\text{SG}} * f_{\text{rev}})(t) + \dot{f}\left(\frac{1}{2\delta}+\frac{\min(t,T_h)}{1-e^{2\delta\min(t,T_h)}}\right), \quad (13)$$

assuming that δ and T_h are known (one can find various methods in the literature to estimate the reverberation time and deduce δ and T_h). The value of $\dot{f}$ can be estimated with a method detailed in [18]. The instantaneous phase of the anechoic signal $\tilde{\varphi}(t)$ is then obtained by integrating $\tilde{f}(t)$ as:

$$\tilde{\varphi}(t) = 2\pi\int_0^t \tilde{f}(\tau)d\tau + \varphi_{\text{rev}}(0). \quad (14)$$

In (14) we have set $\tilde{\varphi}(0) = \varphi_{\text{rev}}(0)$, which is justified by assuming that the signal is preceded by silence, thus the initial phase is unaffected by reverberation.

3. ADAPTATION TO THE STFT FRAMEWORK

The Hilbert transform is not suited to signals made up of multiple sinusoids, this is why we want to adapt the results derived with the Hilbert transform in Section 2 within a time-frequency framework.

3.1. From a full-band to a subband model

Firstly, we have experimentally verified that the convolution product in (5) can be accurately approximated by a convolution product in the Modified Discrete Cosine Transform (MDCT) subbands: $\forall k \in [0, N-1]$, $Y(m,k) \simeq (H(.,k) * S(.,k))(m)$, where $H(m,k)$, $S(m,k)$ and $Y(m,k)$ denote the MDCT of $h(t)$, $s(t)$ and $y(t)$ respectively, at the m-th time frame and the k-th frequency bin.
Then, we have to show that the stochastic model of the RIR in (4) still holds in the MDCT subbands. By using [19] we can indeed compute the MDCT of the product $h(t) = b(t)p(t)$. After straightforward calculations, we get the approximation

$$H(m,k) \simeq \alpha e^{-\delta N m} B(m,k) \quad (15)$$

with $\alpha > 0$ and $B(m,k)$ the MDCT of the white noise $b(t)$. We also know from [19] that $B(m,k)$ is a white noise of same variance σ^2. We then retrieve the stochastic model of the RIR: a white noise damped by a decaying exponential, in each subband of $H(m,k)$.

Finally, by considering the STFT as the analytical part of the MDCT, we can directly apply the results found on the Hilbert phase to the Fourier phase, in each channel of the STFT.

3.2. Estimator with the STFT

The signals are sampled at f_s Hz and the STFT is computed with an analysis window $w(n)$ of size N and a hop size R, resulting in N_f frames and N discrete frequencies $f_k = k\frac{f_s}{N}$, $k \in [0, N-1]$. In Section 2.1 we denoted the Hilbert transform with a hat; from now on we denote the STFT with a hat and an upper case.

To compute the instantaneous frequency of the reverberant signal $f_{\text{rev}}\left(\frac{mR}{f_s}\right)$ at the m-th frame we use the reassigned vocoder framework introduced in [20]. The estimated instantaneous frequency of the anechoic signal $\tilde{f}\left(\frac{mR}{f_s}\right)$ is then computed with (13), dividing N_{SG} by R due to subsampling. Since we assume that each frequency subband contains only one sinusoidal component, we build a two-dimensional array $\tilde{F} \in \mathbb{R}^{N_f \times N}$ carrying the evolution of the estimated frequency in each channel.

For a chirp signal, the phase $\Phi(m,k)$ of its STFT $\hat{S}(m,k)$ is not equal to the Hilbert phase, there is a correction term caused by the analysis window [18]:

$$\Phi(m,k) = \varphi\left(\frac{mR}{f_s}\right) + \arg\left(\Gamma\left(k, f\left(\frac{mR}{f_s}\right)\right)\right) \quad (16)$$

with

$$\Gamma(k,f) = \sum_{n=0}^{N-1} w(n) e^{i\left[2\pi(f-f_k)\frac{n}{f_s}+\pi\dot{f}\left(\frac{n}{f_s}\right)^2\right]}. \quad (17)$$

This is why we construct the estimated anechoic phase $\tilde{\Phi}(m,k)$ at the m-th time frame and frequency bin k with:

$$\tilde{\Phi}(m,k) = \tilde{\Phi}(m-1,k) + 2\pi\tilde{F}(m,k)\frac{R}{f_s} + \arg\left(\Gamma(k,\tilde{f}\,(mR/f_s))\Gamma^{\star}\left(k,\tilde{f}\left((m-1)\frac{R}{f_s}\right)\right)\right) \quad (18)$$

where $.^{\star}$ denotes the complex conjugate. As in (14), we initialize $\tilde{\Phi}(0,k)$ with the reverberant phase.

4. EVALUATION

To test our method we generate anechoic chirps with instantaneous frequencies increasing from 400 Hz to 6 kHz, sampled at $f_s = 16$ kHz. We choose different durations, ranging from $T = 0.5$ s to $T = 3.0$ s. Note that with this frequency sweep, the usual values of $\dot{f}$ one can find in the pitch of human speech would correspond to durations $T \geq 2.0$ s. Each anechoic signal is then convolved with an RIR simulated as in (4), with reverberation times ranging from $\mathrm{RT}_{60} = 0.4$ s to $\mathrm{RT}_{60} = 2.0$ s. Since we only deal with the estimation of the anechoic phase, we use the anechoic magnitude assuming the magnitude spectrogram was perfectly estimated. We synthesize it either with the reverberant phase (no modification), or with the estimated anechoic phase, to generate $s_{\text{ane,rev}}(n)$ or $s_{\text{ane,est}}(n)$ respectively.

Firstly, we focus on frequency estimation and compute for each couple (T, RT_{60}) the instantaneous frequency $f_{\text{ane,est}}(n)$ of $s_{\text{ane,est}}(n)$ and compute the mean relative error:

$$e_{\text{rel}} = \frac{100}{Tf_s}\sum_{n=0}^{Tf_s-1}\left|\frac{f\left(\frac{n}{f_s}\right) - f_{\text{ane,est}}(n)}{f\left(\frac{n}{f_s}\right)}\right|$$

We also compute the mean relative error with $f_{\text{ane,rev}}(n)$ to highlight the effect of the phase modification. As we can see from Figure 2, the relative error with the reverberant phase (dashed lines) increases with $\dot{f}$ and RT_{60}, as expected considering (12). If we look at the relative errors when we use the estimated phase (solid lines) they are approximately divided by a factor 7. Moreover, for usual values of $\dot{f}$ ($T \geq 2.0$ s) the mean relative error is around 2% and is independent of $\dot{f}$ and RT_{60}, which is very satisfying.

Then, we look at the dereverberation provided by the phase modification. To evaluate the dereverberation we use the toolbox released in the last REVERB challenge [21], which computes the Signal-to-Reverberant Ratio (SRR) and the Cepstral Distance (CD) between the clean signal and the enhanced one [11]. We see in Figure 3 that the SRR is increased by 10 dB when using the estimated phase instead of the reverberant one, which is an important gain in term of dereverberation. The CD is also decreased by 5 dB, corresponding to a significant reduction of distortions.

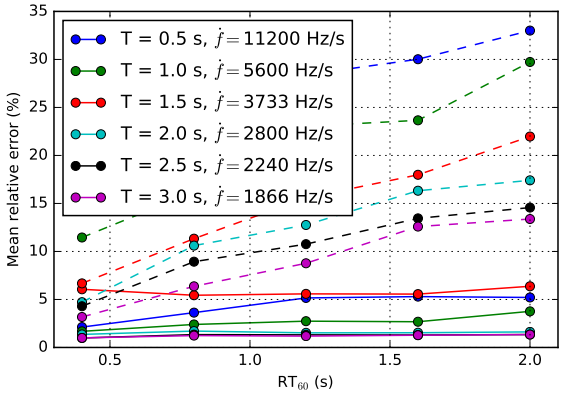

Fig. 2. Errors on the instantaneous frequency estimation

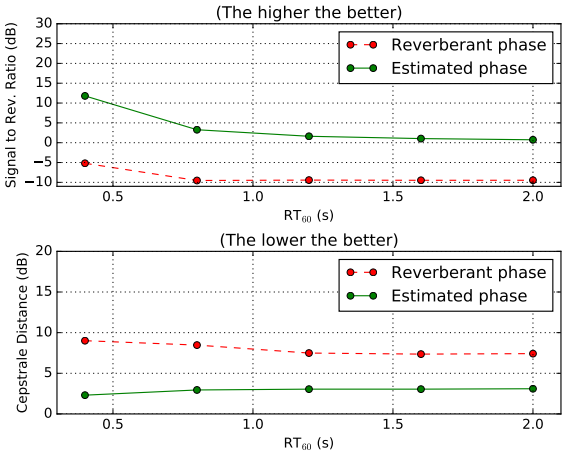

Fig. 3. Evolution of the SRR and the CD for a 3 s chirp

5. CONCLUSION AND FUTURE WORK

By using a stochastic model of RIR, we have derived an expression of the instantaneous frequency of a reverberant chirp. We managed to quantify the influence of the reverberation on this instantaneous frequency, thereby we were able to correct it and estimate the instantaneous phase of the anechoic signal. In order to deal with multicomponent signals, we adapted the estimator to the STFT framework. We tested our method given a set of various reverberant chirps and obtained an accurate frequency estimation. We also significantly improved objective measures of the REVERB challenge, by using the estimated phase instead of the reverberant one.

In future work we will track the instantaneous frequencies of multicomponent signals, with the same reassigned vocoder method [20], in order to estimate the phase of a reverberant signal composed of harmonic chirps. We will then be able to apply our estimator to speech signals, by modeling them as time-varying harmonics + noise [22].

6. REFERENCES

[1] J.Y.C. Wen and P.A. Naylor, "An evaluation measure for reverberant speech using decay tail modelling," in *European Signal Processing Conference (EUSIPCO)*, Florence, Italy, September 2006.

[2] E.A.P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[3] M. Delcroix, T. Hikichi, and M. Miyoshi, "Blind dereverberation algorithm for speech signals based on multichannel linear prediction," *Acoustical Science and Technology*, vol. 26, no. 5, pp. 432–439, 2005.

[4] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 36, no. 2, pp. 145–152, Feb 1988.

[5] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[6] K. Lebart and J.M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *ACUSTICA*, vol. 87, no. 3, pp. 359–366, 2001.

[7] X. Xiao, S. Zhao, D.H. Ha Nguyen, X. Zhong, D.L. Jones, E.S. Chang, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–18, 2016.

[8] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: An insightful benchmark," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 81–85.

[9] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 355–359.

[10] M. Moshirynia, F. Razzazi, and A. Haghbin, "Speech dereverberation method using adaptive sparse dictionary learning," in *Proceedings of REVERB Challenge Workshop*, May 2014, pp. 1–2.

[11] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan 2008.

[12] P.P. Parada, D. Sharma, and P.A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4718–4722.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.

[14] J.D. Polack, *La transmission de l'énergie sonore dans les salles*, Ph.D. thesis, Université du Maine, 1988.

[15] M.R. Schroeder, "New Method of Measuring Reverberation Time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409, 1965.

[16] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, Apr 1992.

[17] A. Savitzky and M.J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[18] M. Betser, *Modélisation sinusoïdale et applications à l'indexation sonore*, Ph.D. thesis, Télécom ParisTech, 2008.

[19] R. Badeau and M.D. Plumbley, "Probabilistic time-frequency source-filter decomposition of non-stationary signals," in *European Signal Processing Conference (EUSIPCO)*, Sept 2013, pp. 1–5.

[20] M. Betser, P. Collen, G. Richard, and B. David, "Estimation of frequency for AM/FM models using the phase vocoder framework," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 505–517, Feb 2008.

[21] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, January 2016.

[22] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonics + noise model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 1993, vol. 2, pp. 550–553.