Estimation of the Spectral Envelope of Voiced Sounds Using a Penalized Likelihood Approach

Marine Campedel-Oudot, Olivier Cappé and Eric Moulines

Abstract

Estimation of the spectral envelope (magnitude of the transfer function) of a filter driven by a periodic signal is a long-standing problem in speech and audio processing. Recently, there has been a renewed interest in this issue in connection with the rapid developments of processing techniques based on sinusoidal modeling. In this paper, we introduce a new performance criterion for spectral envelope fitting which is based on the statistical analysis of the behavior of the empirical sinusoidal magnitude estimates. We further show that penalization is an efficient approach to control the smoothness of the estimation envelope. In low noise situations, the proposed method can be approximated by a two steps weighted least-squares procedure which also provides an interesting insight into the limitations of the previously proposed "discrete cepstrum" approach. A systematic simulation study confirms that the proposed methods perform significantly better than existing ones for high pitched and noisy signals.

Index Terms

Spectral estimation, sinusoidal modeling, speech analysis, non-parametric smoothing

EDICS number: 1-ANLS

I. Introduction

Current speech analysis/synthesis methods capitalize on the source-filter representation of speech signals which is motivated by the acoustic theory of speech production. The basic speech production model which has been proposed more than forty years ago consists of a source signal (glottal excitation) passing through a linear filter (vocal tract) [9], [10]. Depending on the type of speech sound, the excitation signal is either noise-like (unvoiced sounds) or periodic and impulsive (voiced sounds). The filter models several distinct phenomenons (glottal pulse shape, vocal tract transfer function, lips radiation response) and thus does not have a simple and convenient parametric form although the main contribution is that of the vocal tract whose transfer function can be closely approximated by an all-pole filter for most speech sounds. Both because of the properties of the human hearing system and of the signal distortion due to the sound propagating from the speaker to the recording apparatus, only the spectral magnitude of the filter in the source-filter representation is generally thought to be characteristic of the uttered speech sound. In this paper, we

ENST Dpt. TSI / LTCI (CNRS URA 820), 46 rue Barrault, 75634 Paris Cedex 13, France. Fax: + 33 1 45 88 79 35; Email: cappe@tsi.enst.fr.

This work was supported by France Télécom - CNET (Centre National d'Etude des Télécommunications), under contract 95-PE-7697.

specifically consider the estimation of the *spectral envelope* (spectral magnitude of the filter in the source-filter representation) for voiced speech sounds. The reason for focusing on the voiced parts of speech is that for unvoiced sounds, the estimation of the spectral envelope is a classical times-series problem which has been much studied in the parametric case (when assuming for instance that the filter can be modeled by an all-pole transfer function) as well as in the non-parametric one [31], [25]. This is much less true for periodic source signals despite the fact that accurate estimation of the spectral envelope of voiced sounds is a key ingredient in any speech (or audio) analysis/synthesis system which makes uses of voicing decision. These systems are now commonly used for speech coding [20] or speech synthesis and modification [32]. Note that spectral envelope estimation is also of prime importance for speech recognition [28] and although most current speech recognition systems ignore voicing information, pitch and voicing information can be useful even for estimating the spectral envelope, particularly for high-pitched voices.

Early attempts towards identifying the spectral envelope include applying the LPC (Linear Predictive Coding) scheme (which is usually referred to as Auto Regressive, or AR, modeling in the time series literature) in the voiced parts of the signals as well as in the unvoiced ones. This approach performs poorly for high-pitched voiced speech sounds because it is based on the incorrect assumption that the source signal is a second order white noise. Several methods have been designed to overcome this problem in the context of LPC either by further analyzing the LPC residual or by modifying the objective function used for assessing the fit of the AR model [21] [19] [16].

The SEEVOC (Spectral Envelope Estimation VOCoder) technique of [24] is based on the remark that if the source signal is a periodic impulse train, then the observed signal is a sum of sinusoids and thus only provides information concerning the value of the spectral envelope at the frequencies of the harmonics. The solution proposed in [24] consists of interpolating the estimated spectral envelope between these frequency points using a standard smoothing technique. While the SEEVOC approach does not rely on a parametric description of the spectral envelope, authors such as El-Jaroudi and Makhoul [8] and Galas and Rodet [12] have proposed techniques based on the same principle for (respectively) the all-pole and the cepstral representation of the spectral envelope. While the cepstral parameterization may appear to be less justified than the all-pole representation for speech signals, it leads to computationally simpler techniques when the squared log-spectral distance is used to assess the envelop fit [12], [7].

A key point is that we are indeed dealing with an ill-posed inverse problem [22] in trying to recover a whole function (the spectral envelope) from a noisy measurement of its values in a few frequencies (corresponding to the harmonics). Accordingly, in [6], a roughness penalty is added to the envelope fit measure proposed by [12] so as to enforce the smoothness of the estimated envelope, following the so-called "regularization" or "penalization" framework (the latter denomination being more standard in the applied statistics literature). This approach was shown in [6] and [5] to be very efficient in preventing the appearance of unnatural

envelopes observed by Galas and Rodet (usually for high pitched sounds) [12].

The shortcoming of the method described in [6] however is that the use of the squared log-domain distance as a measure of how well the envelope fits the measured harmonic magnitudes is arbitrary and counter-intuitive: Both because the harmonics of high magnitude are more important from a perceptual point of view and because they are more reliably estimated, it would be preferable to give different weights to the fitting errors depending on the magnitude of the harmonics. Although we will not address the first (perceptual) aspect, the point of the present paper consists of showing that the second effect (reliability of the magnitude estimation which depends on the magnitude of the harmonic) can be accounted for using a more elaborate fit criterion. This criterion will be obtained as an approximate likelihood criterion assuming that the ideal voiced speech sound is observed in additive noise. Additive noise is the simplest model which can to some extent account for both: *(1)* the modeling errors (i.e. the fact that the "sum of harmonics" model does not exactly fit a speech signal even on short durations because of the non-stationarity of speech); *(2)* the fact that some voiced sounds also features a significant amount of friction noise; *(3)* the ambient noise which may be of significant level (in mobile communications for instance, Signal-to-Noise Ratios, or in short SNRs, of 5dB or less are not that uncommon).

The rest of the paper is organized as follows: In section II, the approximate likelihood of the envelope parameters is obtained; Section III is devoted to the study of numerical optimization methods suited for maximizing the proposed penalized likelihood criterion; Finally, we discuss in section IV the performance of the method for typical speech analysis purposes.

II. Penalized likelihood criterion

A. Asymptotic integrated likelihood

In voiced parts of the signal, application of the Poisson formula shows that the source-filter representation is equivalent to an harmonic decomposition. We thus assume that the observed signal consists of

$$r_t = \underbrace{\sum_{k=1}^{K} [a_k \cos \omega_k t + b_k \sin \omega_k t]}_{m_t} + \epsilon_t \quad (1 \leq t \leq T) \tag{1}$$

where $\omega_1, \ldots, \omega_K$ are the (radian) frequencies of the harmonics, and ϵ_t is modeled as a (second order) stationary random process with psd. (power spectral density) $\Pi_\epsilon(\omega)$. Note that the fact that the sinusoidal components are harmonics (i.e. that $\omega_k = k\omega_1$) will play no role and that the frequencies $\omega_k, k = 1, \ldots, K$ need not be harmonically related as long as they are well separated in the sense that $\min_{i,j \in \{1,\ldots,K\}} |\omega_i - \omega_j| \gg 2\pi/T$.

Our treatment of the model given by (1) will be based on the assumption that the noise p.s.d. $\Pi_\epsilon(\omega)$ and the component frequencies $\omega_k, k = 1, \ldots, K$ are known. The pertinence

of this assumption for speech processing applications will be discussed in more details in section IV-D. The spectral envelope $S(\omega)$ is parametrized in the power domain such that

$$\begin{aligned} a_k &= \sqrt{S(\omega_k)}\cos\theta_k \\ b_k &= \sqrt{S(\omega_k)}\sin\theta_k \end{aligned}$$

where the phases of the harmonics $\theta_k, k = 1, \ldots, K$ are considered as nuisance parameters.

Let $\hat{a}_k$ and $\hat{b}_k$ denote the amplitude of the phase and quadrature components of the k-th sinusoid estimated from the tapered Fourier transform of the signal:

$$\hat{a}_k = \frac{2}{N_w}\sum_{t=1}^{T} w_t r_t \cos\omega_k t \tag{2}$$

$$\hat{b}_k = \frac{2}{N_w}\sum_{t=1}^{T} w_t r_t \sin\omega_k t \tag{3}$$

where w denotes the data taper (or window), and the normalizing constant N_w is defined as

$$N_w = \sum_{t=1}^{T} w_t$$

To make the expressions simpler, we will assume that the data taper w is obtained by regular sampling of a continuous-time positive window function $\bar{w}(\tau)$ defined on $[0, 1]$, that is: $w_t = \bar{w}(t/T)$ for $t = 1, \ldots, T$.

It is shown in appendix A that in the simpler case where ϵ_t is a Gaussian white noise, the Fourier estimates $(\hat{a}_k, \hat{b}_k), 1 \leq k \leq K$ asymptotically (when T is large) form a set of sufficient statistics for the estimation of the envelope. Moreover, when the nuisance parameters $\theta_k, 1 \leq k \leq K$ are eliminated by marginalization (by integration over the range $[0, 2\pi)$), the resulting asymptotic criterion is a function of the estimated squared magnitudes $x_k = (\hat{a}_k)^2 + (\hat{b}_k)^2$ only. Marginalization is the method of choice for handling nuisance parameters in the Bayesian framework and is generally thought to be more robust than the profile likelihood approach which consists of optimizing with respect to (abbreviated to wrt. in the following) the nuisance parameters [2]. In the case under consideration, using a profile likelihood would imply fitting a complex envelope model to the data. For speech signals however, complex envelope modeling is only a sensible choice if the frame locations can be synchronized with the glottal closures. Such an approach would thus require pitch synchronous processing and robust estimation of the glottal closures, which is a difficult task [11], [3].

The result obtained in appendix A, although restricted in scope, is very intuitive, as it suggests that for large values of the frame size T, estimators of the spectral envelope

should be based on the Fourier power measurements $x_k, 1 \leq k \leq K$ at the frequencies of the harmonics. We will use this principle in a broader context noting that the amplitude values estimated through (2)-(3) are known to be asymptotically normal for a very large class of noise processes (not necessarily white nor Gaussian) and that under suitable technical conditions [14], [27],

$$\mathrm{E}(\hat{a}_k) = a_k + o(1) \quad \mathrm{E}(\hat{b}_k) = b_k + o(1) \tag{4}$$

$$\mathrm{Cov}(\hat{a}_k, \hat{b}_k, \hat{a}_j, \hat{b}_j) = \frac{n_k}{2}\mathbf{I}_4(1 + o(1)) \quad \text{for } k \neq j \tag{5}$$

where $\mathbf{I}_4$ denotes the four dimensional identity matrix, the $o(1)$ notation stands for remainder terms that tend to zero for increasing values of T and n_k is the *apparent noise power* defined as

$$n_k = \frac{4G_{\bar{w}}}{T}\Pi_\epsilon(\omega_k) \tag{6}$$

where $G_{\bar{w}}$ is a normalizing constant which only depends on the type of the analysis window through

$$G_{\bar{w}} = \frac{\displaystyle\int_0^1 \bar{w}^2(u)du}{\left[\displaystyle\int_0^1 \bar{w}(u)du\right]^2}$$

$n_k/2$ corresponds to the power of the noise affecting the measurement of the phase or quadrature amplitude for one sinusoid, and thus decreases in inverse proportion of the sample size [31], [27]. Eqs. (4)-(5) also hold for the maximum likelihood (weighted least-squares) estimator of the sinusoidal amplitudes since both procedures are equivalent for large sample size T [31]. When processing voiced speech with standard analysis settings (frame duration of about 30ms with a smooth data taper), these asymptotic results are indeed accurate because the periodicity of the signal implies that the frequencies of the sinusoidal components are separated by the fundamental frequency which is larger than the spectral resolution, except for the lowest (less than 80Hz) pitch values [31] (note that the results of section IV show that for such very low pitch values, envelope estimation can be reliably achieved by standard methods such as direct AR estimation).

Considering the asymptotic approximation given by (4)-(5), the empiric squared magnitude of the kth harmonic $x_k = (\hat{a}_k)^2 + (\hat{b}_k)^2$ is obtained as the sum of two squared Gaussian variables with non zero means. Up to a scale factor, the resulting variate is distributed according to a non-central χ^2 distribution with two degrees of freedom, or Rice distribution, whose probability density is given by [17]

$$p(x_k) = \frac{1}{n_k}\ \exp[-\frac{s_k + x_k}{n_k}]\ I_0\left(2\sqrt{\frac{s_k x_k}{n_k^2}}\right) \tag{7}$$

where $s_k = a_k^2 + b_k^2$ is the actual value of squared magnitude of the kth harmonic and $I_0(\cdot)$ stands for the modified Bessel function of the first kind and order $\nu = 0$ [1]:

$$I_\nu(y) = \frac{1}{\pi} \int_0^\pi e^{y \cos\theta} \cos(\nu\theta) d\theta \tag{8}$$

Note that (7) could also be expressed in terms of the series expansion of $I_0(\cdot)$ as [1], [17]:

$$I_\nu(y) = (\frac{1}{2}y)^\nu \sum_{p=0}^{+\infty} \frac{y^{2p}}{2^{2p} p!(p+\nu)!} \tag{9}$$

Eq. (7) corresponds to a positively skewed distribution (especially for low values of s_k since x_k is by construction positive) with mean and variance [17]:

$$\begin{cases} \mathrm{E}(x_k) &= s_k + n_k \\ \mathrm{Var}(x_k) &= n_k(2s_k + n_k) \end{cases} \tag{10}$$

With the independence approximation, the negative log-likelihood of the K squared amplitude estimates $L(x_1, \ldots, x_K|S)$ may be written as

$$L(x_1, \ldots, x_K|S) = \sum_{k=1}^{K} \left[\log n_k + \frac{s_k + x_k}{n_k} - \log I_0 \left(2\sqrt{\frac{s_k x_k}{n_k^2}} \right) \right] \tag{11}$$

In appendix A, the preceding expression is obtained directly using simple calculations for the case of Gaussian white noise.

In practice, direct evaluation of (11) using (9) can be awkward because the Bessel functions have an exponential behavior in $+\infty$. The computation of $\log(I_0)$ (as well as I_1/I_0, introduced in appendix B) can however be carried out using standard combinations of series truncation and approximations detailed in [1], [17] or [26].

B. Roughness Penalty

In many cases of interest, direct minimization of (11) yields envelopes that have a non-smooth behavior and are unacceptably sensitive to small variations in the observed data [6], [25]. This phenomenon has been previously observed with other envelope estimation methods [8], [12]. Intuitively, the ill-posed character of the envelope estimation problem is a consequence of the fact that there are many continuous envelopes that can be plausibly fitted to just one snapshot of a reduced set of frequency measurements.

The standard solution to this problem consists of constraining the behavior of the estimated envelope by use of a so-called *roughness penalty* $R(S)$ (also known as a "regularization" or "smoothing" functional) [22]. The likelihood criterion is replaced by a penalized criterion of the form:

$$L(x_1, \ldots, x_K|S) + \lambda R(S) \tag{12}$$

where $R(S)$ is the roughness penalty which takes large values for envelopes S that have a non-smooth behavior and λ is a scalar parameter which controls the smoothness of the estimated envelope. This penalized approach can also be viewed as a Bayesian maximum a posteriori estimation procedure where $\exp(-\lambda R(S))$ plays the role of a prior for the envelope parameters [22].

In the following, we use

$$R(S) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{d^r \log S(\omega)}{d\omega^r} \right]^2 d\omega \tag{13}$$

which has been used for envelope estimation in [6] (with $r = 1$) and in [25] (with $r = 2$). Note that (13) features the derivative of $\log S(\omega)$ rather than that of $S(\omega)$. This choice is motivated by two reasons: The log-scale is generally considered to be perceptually more meaningful for speech spectra than the linear scale and this choice makes the effective computation of $R(S)$ much simpler (see section III). It is usually found, and this is also true for the problem under consideration, that the choice of the roughness penalty has less influence than the value of the smoothing parameter λ [22], [15]. In the following, we use (13) with $r = 1$ and we postpone the discussion of the influence of λ to section IV-B.

III. Cepstral envelope estimation algorithms

The penalized likelihood criterion given by (12) may be used in various ways. Because it derives from a likelihood approximation, it is robust to the envelope parameterization and we have successfully applied the method to both the cepstral and the all pole envelope representations. Another interesting feature of (12) is that it is naturally compatible with other forms of likelihood approximation, such as the "Whittle likelihood" used for stationary processes with smooth psd. [25]. In [4], [23] this property is used to estimate a single spectral envelope in a two-band model where the lower part of the spectrum is modeled as an harmonic signal (for voiced sounds) and the upper part of the spectrum is modeled as a stationary process.

After a bit of experimentation with the method, the approach we recommend consists of optimizing (12) using the cepstral parameterization. Indeed, the penalized criterion of (12) does not in general correspond to a convex function and its minimization has to be carried out using an iterative numerical optimization procedure. The optimization turns out to be much faster and reliable (i.e. free of local extrema) when using the cepstral parameterization. One first reason for this good behavior is that the penalty $R(S)$ is then a (convex) quadratic form [6], [25] (see also section III-A below); As a second element to support this finding, we show in appendix B that the likelihood criterion given by (11) is convex with high probability in a neighborhood of the true envelope for low noise levels.

Finally, because there are speech processing applications for which numerical optimization would be too demanding, we discuss in section III-B a low implementation cost approx-

imation of (12) under the form of an equivalent (for low noise levels) weighted least-squares criterion.

A. Exact algorithm

In this section, we consider the form taken by the proposed procedure when using the cepstral parameterization of the envelope S:

$$S(\omega_k) = s_k = \exp[c_0 + 2\sum_{n=1}^{p} c_n \cos \omega_k n] \tag{14}$$

With this parameterization, application of the Parseval relation to (13) shows that $R(S)$ reduces to a quadratic form:

$$R(S) = \mathbf{c}'\mathbf{R}\mathbf{c} \tag{15}$$

where $\mathbf{c} = (c_0, \ldots, c_p)'$ is the vector of cepstrum coefficients (the prime denoting transposition), and $\mathbf{R}$ is a diagonal matrix whose diagonal entries are $2(0, 1^{2r}, 2^{2r}, \ldots, p^{2r})$ [6], [5]. Thus, for the cepstral parameterization, $\exp(-\lambda R(S))$ exactly corresponds to a multivariate normal prior, with zero mean and variance decreasing with the cepstrum index n proportionally to $1/n^2$ (when $r = 1$), which resembles the observed statistical behavior of speech cepstrums[1] [18].

With the cepstral parameterization, exact computation of the gradient of the composite criterion given in (12) is feasible: Eq (15) shows that the gradient of the penalty $R(S)$ is given by $2\mathbf{R}\mathbf{c}$ and the gradient of the integrated likelihood criterion may be computed as (see appendix B)

$$\boldsymbol{\nabla}_{\mathbf{c}} L(\mathbf{c}) = \sum_{k=1}^{K} \begin{pmatrix} 1 \\ 2\cos\omega_k 1 \\ \vdots \\ 2\cos\omega_k p \end{pmatrix} \frac{s_k}{n_k} \left[1 - \sqrt{\frac{x_k}{s_k}} \frac{I_1}{I_0}\left(2\sqrt{\frac{s_k x_k}{n_k^2}}\right)\right] \tag{16}$$

It is thus possible to use efficient iterative optimization approaches for minimizing (12). In the following, we use a BFGS quasi-Newton method with embedded cubic polynomial line searches [26] to estimate the envelope parameters. This is not necessarily the best available method for unconstrained numerical optimization but it is implemented by most numerical analysis packages so that our results will be easily reproduced.

Experimenting with different initializations suggests that the optimization procedure is not very sensitive to its initialization (or in other words that local minima are not a real problem), except when the value of λ is too low (see section IV-B). Appendix B provides an element of the answer by showing that *when the cepstral parameterization is used*, the

[1] This is actually true only for the sub-vector $c_1, \ldots, c_n$ since $R(S)$ does not depend on c_0. For c_0 the prior equivalent to $R(S)$ is thus an improper constant prior, independent of $c_1, \ldots, c_n$.

integrated likelihood criterion $L(\cdot)$ is convex with high probability in a large neighborhood of the actual envelope. Note that the inclusion of the roughness penalty further amplifies this effect by adding a constant positive definite matrix ($\mathbf{R}$) to the Hessian of the criterion to be minimized.

B. Weighted least-squares approximation

Depending on the constraints of the application under consideration (computing resources and floating point precision) the iterative optimization approach described in the previous section may be too demanding. For this purpose, we now derive an approximation of (7) based on the delta or approximate linearization method which is suitable for low noise conditions. Starting from the joint asymptotic normality of the phase and quadrature estimates $\hat{a}_k$ and $\hat{b}_k$, we obtain the asymptotic normality of the transformation $v_k = \log(\hat{a}_k^2 + \hat{b}_k^2)$ using standard arguments [34], where the limiting mean and covariance are respectively given by

$$\begin{aligned} \mathrm{E}(v_k) &= \log(a_k^2 + b_k^2) + o(1) \\ &= \log s_k + o(1) \\ \mathrm{Var}(v_k) &= \left(\frac{\partial v_k}{\partial a_k}, \frac{\partial v_k}{\partial b_k}\right) \mathrm{Cov}(\hat{a}_k, \hat{b}_k) \left(\frac{\partial v_k}{\partial a_k}, \frac{\partial v_k}{\partial b_k}\right)' (1 + o(1)) \\ &= 2\frac{n_k}{s_k}(1 + o(1)) \end{aligned} \tag{17}$$

and in addition, one can also show by the same technique that (5) implies that v_k and v_j jointly are asymptotically normal and that

$$\mathrm{Cov}(v_k, v_j)/\sqrt{n_k n_j} \to 0 \quad \text{when } k \neq j$$

If we assume $s_k, 1 \leq k \leq K$ to be the actual values of the envelope at the harmonic frequencies, the optimally weighted least squares criterion is thus given by

$$\sum_{k=1}^{K} \frac{s_k}{n_k} \left(\log x_k - \log s_k\right)^2 \tag{18}$$

Eq. (18) is close to the discrete cepstrum criterion proposed by Galas and Rodet in [12] with the important difference that instead of giving equal weights to all the frequency measurements when doing the least-squares fit, one should weight them according to the local SNRs s_k/n_k. The pertinence of this weighting scheme is dramatically illustrated by fig. 1 which shows how the reliability of the estimated amplitudes decreases in noisy area of the spectrum.

Eq. (18) shows that the optimal choice for the least-squares weights depends on the local SNRs and thus on the unknown spectral envelope S. To approach this optimal behavior with

a data driven approach (that is without requiring prior information on the envelope to be estimated), we adopt a classic approach in non-parametric smoothing based on a preliminary estimate of S (sometimes referred to as a "plug-in" approach):

Algorithm: Cepstral estimation based on the Gaussian approximation

1. Compute the penalized least-squares solution

$$\hat{\mathbf{c}} = \left(\mathbf{C}'\mathbf{C} + \lambda\mathbf{R}\right)^{-1}\mathbf{C}'\mathbf{v} \tag{19}$$

where $\mathbf{C}$ is the cepstrum regression matrix

$$\mathbf{C} = \begin{pmatrix} 1 & 2\cos(\omega_1) & \dots & 2\cos(\omega_1 p) \\ \vdots & \vdots & & \vdots \\ 1 & 2\cos(\omega_L) & \dots & 2\cos(\omega_L p) \end{pmatrix} \tag{20}$$

$\mathbf{R}$ is the regularization matrix defined in (15) and $\mathbf{v} = (\log x_1, \dots, \log x_K)'$ is the vector of log-power measurements.

2. Compute the weights $\gamma_k = \hat{s}_k/n_k$ for $k = 1, \dots, K$ where $\hat{s}_k$ is computed using (14) from the vector of cepstral coefficients $\hat{\mathbf{c}}$ estimated using (19).
3. Solve the penalized weighted least-squares problem by

$$\hat{\mathbf{c}} = \left(\mathbf{C}'\mathbf{\Gamma}\mathbf{C} + \lambda\mathbf{R}\right)^{-1}\mathbf{C}'\mathbf{\Gamma}\mathbf{v}$$

where $\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \dots, \gamma_K)$ is the diagonal matrix of weights.

The numerical complexity of the above algorithm is twice that of the discrete cepstrum method (step 1 only, with $\lambda = 0$), whose numerical complexity is of order p^3 (solution of a linear system with $p+1$ unknowns). In the following, we will use the phrase "least-squares" to refer to the method of Galas and Rodet rather than "discrete cepstrum" which is potentially misleading in a context where several cepstral estimation algorithms are compared.

IV. Evaluation

In this section, we discuss the performances of the various estimation methods introduced in section III which are referred to as: **OLC** for "Optimization of the Likelihood Criterion" (cf. section III-A), **LS** for "Least-Squares" approximation and **WLS** for "Weighted Least-Squares" approximation (cf. section III-B). For comparison purposes, the performances of the standard Auto-Regressive (or **AR**) approach on the same data are also reported.

The experimental setup is first described in section IV-A. After investigating the influence of the smoothing parameter λ (section IV-B), we then compare the performance of the four methods (section IV-C). Finally, section IV-D is devoted to robustness issues and to the application of the method in a complete harmonic analysis/synthesis system.

A. *Experimental setup*

For real speech signals, there is no way of controlling what the actual envelope is and furthermore, the degree of adequacy of the sinusoidal model itself is difficult to assess. To base our analysis on objective distance measures, we thus consider synthetic signals generated from the model given in (1) for three typical speech envelopes and various pitch and SNR combinations summarized in table I (see section IV-D for an example of results obtained on real speech).

Parameter	Values
Envelope	/a/ (E1), /u/ (E2), /i/ (E3)
Pitch ($4000 \times \omega_1/\pi$)	100, 140, 180, 220, 260Hz
Signal-To-Noise-Ratio (SNR)	50, 40, 30, 20dB

TABLE I
SUMMARY OF THE SIMULATION PARAMETERS.

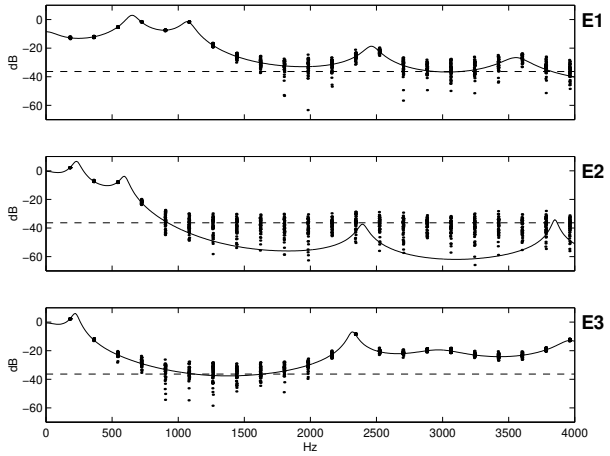

Fig. 1. Scatterplot of the estimated harmonic magnitudes for 50 independent noise realizations for each of the three test envelopes; The solid curve is the actual envelope and the dotted line corresponds to the apparent noise level n (SNR = 20dB, pitch 180Hz).

The envelopes have been obtained by convolving four cascaded second order cells designed from formant data with a stylized glottal pulse shape and lip radiation model, following [29]. To avoid aliasing effects due to the high frequency formants, the computation of the envelope was done using 16kHz as sampling frequency, retaining only the first half of the spectrum. The three resulting envelopes are shown on figure 1 (solid line). The envelopes are used to generate 8kHz sampled synthetic signals according to (1), which sound reasonably natural and are clearly recognizable as vowel sounds [a], [u] and [i]. The duration of the frame is set to $T = 256$ samples (32ms). In all experiments, $\bar{w}$ is chosen to be a hanning window. To assess the robustness to noise of the various estimation schemes, 50 independent

noise realizations where generated for each of the 60 parameter combinations arising from table I. For simplicity we use Gaussian white noise so that the apparent noise power defined in (6) is constant through the frequency domain.

Figure 1 shows the variability of the Fourier magnitude estimates x_k for the middle pitch (180Hz) and high noise (20dB SNR) condition. As a first remark, we note that the second envelope has a much important spectral dynamic so that it is indeed much more sensitive to noise than the other two. The second important remark is that figure 1 clearly highlights the weakness of the LS criterion which treats all the measured Fourier magnitudes as being equally credible whereas it is patent that the measurements in regions where the envelope level is close (or below) the apparent noise level are not at all reliable.

As a reference, an AR(12) model was fitted to all signals using Yule-Walker method with windowing by the hanning window. The corresponding AR envelope estimate was obtained as

$$\mathcal{E} = \frac{\hat{\sigma}^2}{|\hat{A}(\mathrm{e}^{-j\omega})|^2} \times \frac{8}{3} \times \frac{2}{K} \tag{21}$$

where $8/3$ corresponds to the inverse of $\int_0^1 \bar{w}^2(u)du$ for the hanning window, which is the power correction due to windowing, and $2/K$ is a scaling factor which takes into account that a value of 1 for the spectral envelope corresponds to a sum of K sinusoids which is a signal of power $K/2$ while we assume when deriving the AR estimate that the input is a white noise of power 1. Note that because we are considering different pitch frequencies ω_1, the number of harmonics $K = \lfloor \pi/\omega_1 \rfloor$ varies significantly. This scaling procedure is an ad hoc approach to compensate for the fact that we are considering an incorrect model of the data when fitting an AR model. It nonetheless gives satisfying results in regions of the spectrum that are free from noise (as in figure 6-A1). This method was selected so as to minimize the average envelope estimation error on the simulated data for the 50dB SNR condition among a number of alternatives which included: use of a fixed scaling factor determined from the data or scaling according to (21); hanning windowing or no windowing; AR orders between 8 and 16. Note that all options had a limited influence except for the model order.

The distance between the actual envelope S and an estimated envelope $\hat{S}$ is computed as a discrete approximation to

$$\sqrt{\frac{1}{\nu_{\max} - \nu_{\min}} \int_{\nu_{\min}}^{\nu_{\max}} \left(10 \log_{10} S(\nu) - 10 \log_{10} \hat{S}(\nu)\right)^2 d\nu} \tag{22}$$

where the νs correspond to the frequencies between $80Hz$ ($\nu_{\min}$) and $4kHz$ ($\nu_{\max}$) expressed on the Bark (critical band rate) scale using the non-linear frequency warping function given by [35]. This distance measure (expressed in dB) is thus exactly the one used for speech recognition applications [28] and is generally thought to be perceptually more significant than the log-spectral RMS error computed on the original frequency scale. In the present

experiment, the Bark transformation mostly has the effect of slightly compressing the error values, and thus reduces the measured differences between the estimation methods. Note that warping the harmonic frequencies on the Bark frequency scale can improve the performances of the methods of section III wrt. the criterion given in (22) as demonstrated in [5] (this possibility is not considered here for reasons of space).

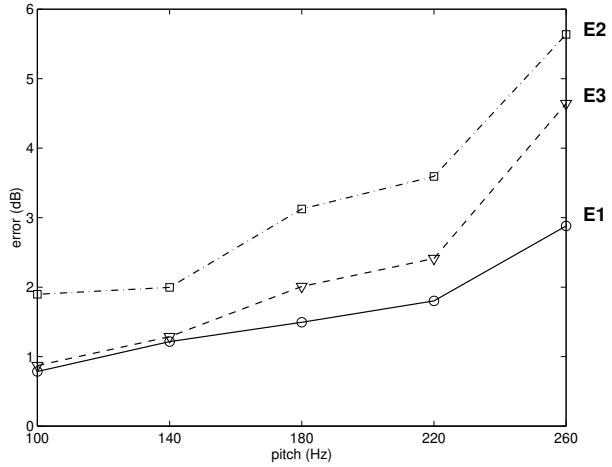

Fig. 2. Median AR estimation error as a function of pitch for the three envelopes (50dB SNR).

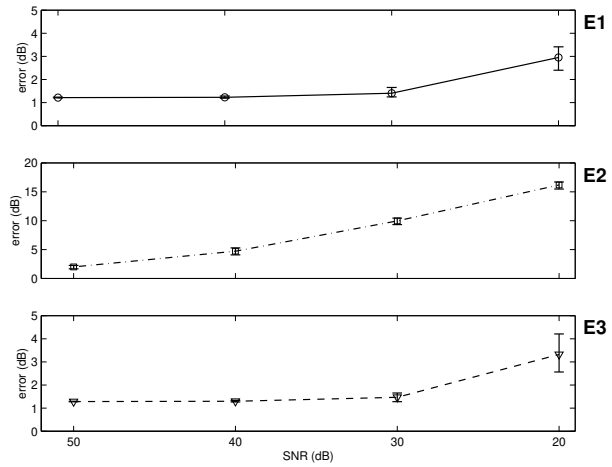

Fig. 3. Median AR estimation error with 5 and 95% quantiles, as a function of the SNR for the three envelopes (140Hz pitch).

Figures 2 and 3 show the two main factors which influence the AR estimation method: The first one is the pitch frequency with results worsening steadily as the pitch raises (fig. 2). This well known effect illustrated by figure 6 (compare the A1 and A2 plots), occurs when the harmonic frequencies are sufficiently spaced apart and manifests itself by a bias of the envelope resonances which are attracted by the nearest harmonic frequency.
The second factor of influence is the noise which significantly degrades the performances

of the method. Figure 7 shows that in noisy areas of the spectrum, the estimated envelope grossly overestimates the actual envelope by fitting the noise spectrum rather than the signal spectral envelope. Because of the normalization given by (21), the envelope estimate in noisy areas of the spectrum is located above the apparent noise level (dotted line in figure 7) and the AR behavior in these regions is close to that of the methods based on "peak picking" from the periodogram.
In both figures 2 and 3, the error values pertaining to the E2 envelope are much larger than those corresponding to the other two envelopes because of its important spectral dynamic.

B. Influence of the smoothing parameter

A potential problem with the methods discussed in section III is that they involve a smoothing parameter λ which could be difficult to tune properly. To investigate this problem, all the signals from the simulation database where analyzed using 50 different values of the smoothing parameter λ logarithmically spaced between 10^{-3} and 10. There are two different ways of constraining the estimated cepstral envelopes to be smooth (this is valid for the three – LS, WLS and OLC – methods): one consists of reducing the order p of the cepstral decomposition in (14), and the other consists of increasing the value of the smoothing parameter λ. It turns out that the second one is the most effective since selecting a cepstral order p which is too small can generate for some envelopes a large unrecoverable approximation bias. In the following, we thus use $p = 40$, that is more parameters than the number of harmonics (for all pitch values) so that the smoothness of the envelope is fully determined by λ. Of course the methods can be used with success for reduced order cepstral parameterization (typically, of the order of twenty coefficients are needed to obtain a reasonable approximation to envelopes such as the ones shown on figure 1) but then, the value chosen for λ should also depend on the choice of p.

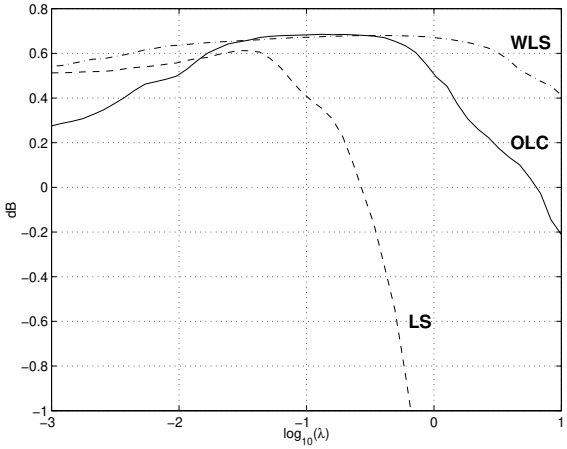

Fig. 4. Median reduction in estimation error wrt. the AR method as a function of λ for LS, WLS and OLC.

Figure 4 shows the median *reduction in estimation error wrt. the AR method*, ie.

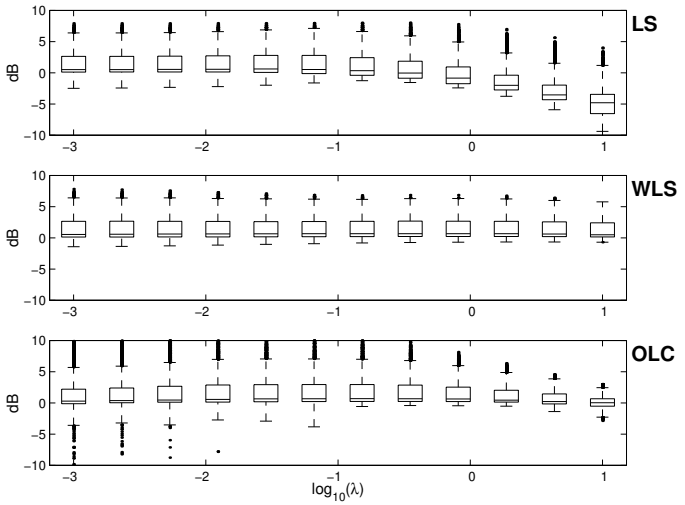

Fig. 5. Box and whisker plots of the reduction in estimation error wrt. AR estimation, for 12 values of λ.

SNR (dB)	50	40	30	20
Median loss (dB)	0	0	0	0.3
Upper 90% quantile (dB)	0	0.5	1.7	3.9

TABLE II
MEDIAN REDUCTION IN ESTIMATION ERROR OF OLC WRT. WLS AS A FUNCTION OF THE SNR ($\lambda_{\mathrm{WLS}} = 0.6$, $\lambda_{\mathrm{OLC}} = 0.15$).

$\mathcal{E}_{\mathrm{AR}} - \mathcal{E}_{\text{other method}}$ considering all the signals in the database (60 conditions times 50 noise realizations), for the three methods. Comparing to the performances of the AR method on the same signals reduces the variability and ensures that we are focusing on the improvements and not on the absolute values of the error which vary to a great extent with the envelope, the noise condition, etc (see figures 2 and 3). A first important remark is that there are large ranges for λ (allowing for variations of several orders of magnitude) where the median reduction in error is positive, that is where the three methods perform better than AR. As suggested by figure 1, the LS method is less efficient than both WLS and OLC, and more sensitive to the choice of λ, with performances degrading quickly in the rightmost part of the plot.

To give an idea of the implementation costs of OLC, the median number of iterations defined as the number of evaluations of the criterion (11) and its gradient (16) is 67, and in 50% of the cases the required number of iterations is between 55 and 80. Note that because we are using a quasi Newton approach, the optimization converges quite quickly once it has reached the domain of attraction of a mode [26], so that these numbers are rather independent of the selected stopping criterion.

Figure 5 shows a more detailed picture by plotting the distributions of the reduction

in estimation error for several values of λ under the form of box and whisker plots (with the box showing the median and the 25 and 75% quantiles, the whiskers giving an idea of the extent of the distribution and the points indicating "outliers"). For WLS and OLC, the distributions of the error reduction are almost entirely located above 0 which indicate that the improvement wrt. AR is quasi-systematic (and not only true on average). The OLC plot (bottom plot in figure 5) shows two interesting facts:
First, for low values of λ, the OLC performs very badly for some rare envelopes (outliers falling below -5dB for the values of λ smaller than 0.1). These cases indeed correspond to situations of misconvergence of the method where the upper limit of 250 iterations is reached without stabilization of the envelope estimate. The fact that these cases of misconvergence only occur when λ is too small is coherent with the discussion of section III concerning the role of the roughness penalty.
Although WLS performs nicely and is most robust to the choice of λ, there are cases where the OLC error is much lower (upper outliers in the bottom plot of figure 5). Indeed, the OLC method is more robust to noise than WLS as illustrated by table II: Whereas the two methods are absolutely equivalent when the noise is as low as 50dB SNR, OLC does significantly better for the 20dB SNR condition with an improvement that is greater than 3.9dB in 10% of the cases. An example of the difference of performances between WLS and OLC in noisy situations will be given below in figure 7.

Based on figure 4, the optimal choice for a fixed value of λ is $3.5\,10^{-2}$ for LS, 0.6 for WLS and 0.15 for OLC. The potential gain of tuning λ for each signal separately is rather weak as the previous choices ensure median performances that are less than 0.05dB from optimal for the three methods (the optimal choice of λ being in this case computed independently for each signal). It is only for OLC that data dependent tuning of λ could be of some interest, since there is a few cases where the loss wrt. the optimal choice of λ is significant: 12% of cases where it is greater than 0.5dB and 5% where it is greater 1.5dB. Unfortunately data driven tuning of the smoothing parameter for the criterion given by (12) is an open problem because the envelope depends non-linearly on the parameters and furthermore has an influence on the hypothesized noise level (as illustrated by figure 1). In the following, we thus only consider the performances of the method obtained when setting λ to the fixed values given above, which seems the most reasonable option for speech and audio processing.

C. Detailed analysis of the performances

The first type of situation where OLC or WLS are superior to AR is when the pitch frequency is high. Table III shows that both methods perform equally well for the lower pitch conditions (100 and 140Hz), but OLC is preferable when the pitch is high, with a difference which can be as high as 2.3dB on average for the 260Hz pitch. Figure 6 shows a typical example of this situation where both methods are equivalent for the 100Hz pitch (A1 and B1 plots) and the AR method is severely biased towards the frequencies of the harmonics located at 660Hz and 1.1kHz when the pitch equals 220Hz (A2 and B2 plots). The results

Pitch (Hz)	100	140	180	220	260
Lower 10% quantile (dB)	0.1	-0.2	0.1	-0.1	0.3
Median reduction (dB)	0.3	0.1	0.5	0.4	2.3
Upper 90% quantile (dB)	0.4	0.3	0.7	0.7	3

TABLE III

INFLUENCE OF PITCH: MEDIAN REDUCTION IN ESTIMATION ERROR OF OLC WRT. AR AS A FUNCTION OF THE PITCH (50DB SNR).

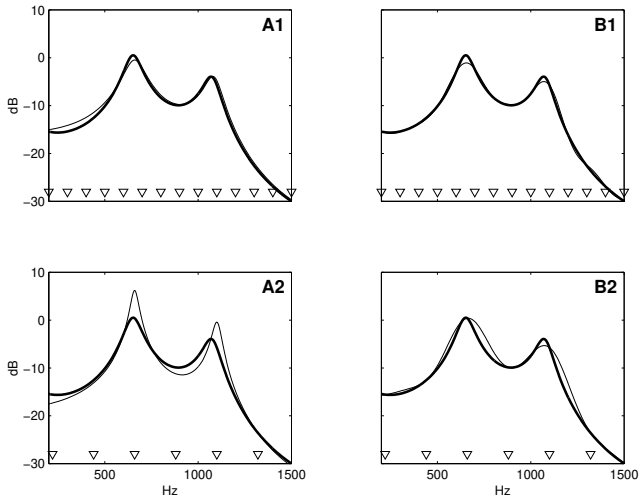

Fig. 6. Influence of the pitch frequency (50dB SNR) : Estimated envelope (light curve) and actual E1 envelope (bold curve) in the 200Hz-1.5kHz band. **A1** AR method with 100Hz pitch; **B1** OLC method with 100Hz pitch; **A2** AR method with 220Hz pitch; **B2** OLC method with 220Hz pitch. The triangles represent the frequencies of the harmonics.

of WLS are not represented on table III and figure 6 as these pertain to the 50dB SNR condition for which the WLS and OLC estimates are indistinguishable (cf. table II).

The presence of noise is the other situation where OLC and WLS are more accurate than AR. Table IV shows the difference between AR an OLC becoming quickly significant as the SNR decreases. Figure 7, which shows three superimposed envelope estimates in each plot to give an idea of the variability, illustrates the origin of the measured differences: While regions where the envelope lies well below the noise level cannot be estimated precisely by any of the methods, OLC and WLS largely reduce the envelope over-estimation effect as well as the variability caused by the noise. In such a situation, OLC performs better than WLS, which is not surprising since the Gaussian approximation used to derive the WLS criterion (in section III-B) is very poor in noisy regions of the spectrum.

SNR (dB)	50	40	30	20
Lower 10% quantile (dB)	0.1	0	0.1	1.2
Median reduction (dB)	0.3	0.3	0.4	2.4
Upper 90% quantile (dB)	0.4	2.6	5.2	8.7

TABLE IV
INFLUENCE OF NOISE: MEDIAN REDUCTION IN ESTIMATION ERROR OF OLC WRT. AR AS A FUNCTION OF THE SNR (100HZ PITCH).

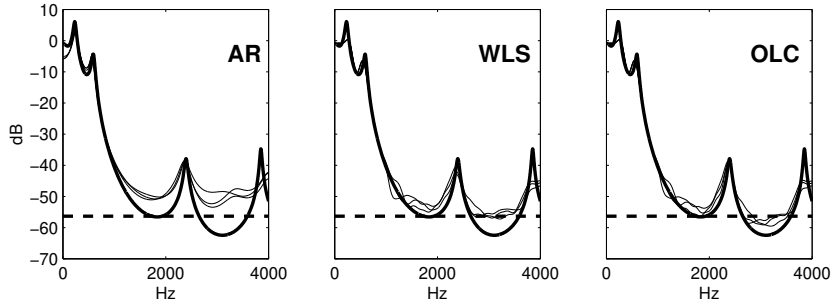

Fig. 7. Envelopes estimated by AR, WLS and OLC (from left to right). Each plot consists of the actual envelope (bold curve), envelope estimates for three noise realizations (light curves) and the apparent noise level (dotted line). E2 envelope, 140Hz pitch, 40dB SNR.

D. Robustness issues

Until now, the simulation parameters (pitch and noise level) have been considered as known, which favors the methods which make use of this information – LS, WLS and OLC for the pitch, WLS and OLC for the noise level.

The methods proposed in this paper are certainly sensitive to pitch estimation errors, but sinusoidal (or harmonic) modeling is by definition very vulnerable to pitch errors: Local pitch estimation is indeed very reliable (see [27] for details) so that when errors actually occurs, they are usually quite "large" (jump to a sub-multiple, incorrect voicing decision). Such an error is much more audible than isolated envelope estimation errors. A reliable pitch detector is thus an absolute requirement for sinusoidal modeling. From our experience, the most troublesome points are incorrect estimation of the noise level and/or incorrect voicing decisions. It is indeed well known that the fact that the stationary model (1) does not exactly fit the signal, even in voiced sections, makes estimation of the noise psd. and assessment of the fit of the harmonic model a difficult issue. Recall that the "noise" corresponds to anything that is not fitted by the harmonic model. In practice, the "noise" thus corresponds both to speech related sounds (friction noise for instance) and/or to environmental sounds. Note that incorrect voicing decisions are also a problem for systems that use AR envelope estimation: Figure 7 clearly shows that an AR method, modified so as to produce an estimate of the envelope in voiced parts of the spectrum, is a biased estimate of the psd. in noisy

area of the spectrum and vice-versa. It is nonetheless patent that as AR estimation does not require estimating the noise psd. it is more robust than WLS and OLC in this respect.

Pitch (Hz)	100	140	180	220	260
10dB underestimation of the noise level					
Median reduction (dB)	0.2	0	0.5	0.3	2.2
10dB overestimation of the noise level					
Median reduction (dB)	0.3	0.1	0.5	0.2	2.2

TABLE V

INFLUENCE OF PITCH WHEN THE NOISE LEVEL IS UNDER/OVER ESTIMATED: MEDIAN REDUCTION IN ESTIMATION ERROR OF OLC WRT. AR AS A FUNCTION OF THE PITCH (50DB SNR).

SNR (dB)	50	40	30	20
10dB underestimation of the noise level				
Median reduction (dB)	0.1	0.2	0.2	1.4
Upper 90% quantile (dB)	0.3	1.8	3	3.5
10dB overestimation of the noise level				
Median reduction (dB)	0.3	0.2	0.4	2.2
Upper 90% quantile (dB)	0.5	0.3	5.9	7.3

TABLE VI

INFLUENCE OF NOISE WHEN THE NOISE LEVEL IS UNDER/OVER ESTIMATED: MEDIAN REDUCTION IN ESTIMATION ERROR OF OLC WRT. AR AS A FUNCTION OF THE SNR (100HZ PITCH).

In order to provide some quantitative elements to the above discussion, we evaluated the analogous of tables III and IV with a systematic mis-estimation of $\pm$10dB of the noise level. Table V shows that, as expected, the effect of noise under/over-estimation is not significant when the noise level is small. In noisy situations, comparison of table VI with table IV shows that the situation is more contrasted with a limited impact of noise overestimation and a more significant degradation in case of underestimation. Overestimation indeed means treating as dubious some measurements that are already affected by noise whereas underestimation can constrain the envelope to take into account measurements that are mostly dominated by noise (cf. fig. 1). In both cases, the median reduction in error stays positive which means that OLC is still preferable to AR despite a severe error in the estimation of the actual noise level.

To illustrate the robustness of the proposed approach, we now consider the more realistic case where all model parameters are unknown and need to be estimated. The analyzed signal now is a 0.87s voiced section of good quality real speech[2] uttered by a young children. The

[2] IPA transcription: [amɑ̃ʒele]

pitch is quite high and varies significantly over the selected excerpt (from 265Hz in the middle section to 480Hz at the end of the excerpt). The signal is analyzed with 30ms frames shifted by 5ms. The pitch is determined using the method of [13] without frame-to-frame pitch tracking. To estimate the noise psd, we follow the suggestion of [30] and discard the periodogram values corresponding to frequency indexes located near the harmonic frequencies. A smooth psd. model is then fitted to the remaining periodogram ordinates using the approach of [25]. The time-domain synthetic signal is obtained by overlap-add using the estimated phase of the harmonics together with the harmonic amplitudes computed from the estimated spectral envelope.

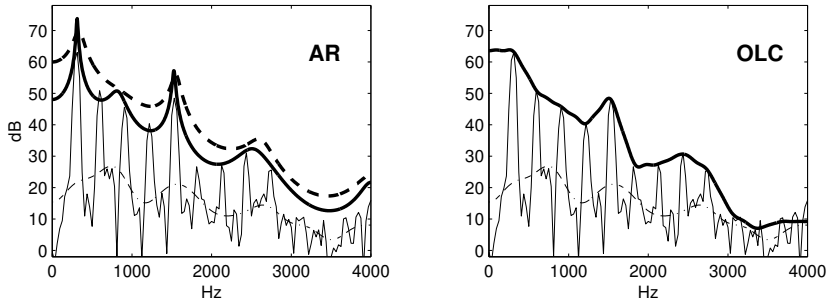

Fig. 8. Estimation results for the frame located 1.35s from the beginning. Left: AR (bold curve) and smoothed AR envelope (dashed bold curve); Right: OLC envelope (bold curve). On both plots: squared magnitude spectrum scaled by $4/N_w^2$ (light curve) and estimated noise psd. scaled by $4G_{\bar{w}}/T$ (dotted curve).

Comparing the bold curves on the two plots of figure 8 clearly illustrates two shortcomings of the AR approach in this context: Ringing (because the pitch is very high, some of the poles are located exactly at the harmonic frequencies) and overestimation in noisy areas of the spectrum. By contrast, the envelope estimated by OLC is both smoother and more precise and, in the upper region of the spectrum (above 3kHz) where the harmonics are dominated by noise, it is less influenced by noise than the AR envelope. Several modifications of AR modeling have been suggested in order to circumvent the ringing problem. As an illustration of this type of approaches, the dashed bold curve in the left plot of figure 8 corresponds to the result obtained with the spectral smoothing technique of [33] (which consists in weighting the estimated autocovariance coefficients). The smoothed AR estimation is not a very attractive technique compared to OLC in this situation because the elimination of the ringing phenomenon is obtained at the cost of a significant overestimation of the envelope in the valleys (in addition, the use of AR smoothing for low pitches where ringing does not occur is not recommended since it severely degrades the accuracy of AR envelope estimation).

The spectrograms shown in figure 9 convey the same idea with AR envelope estimation resulting in patent overestimation of the magnitude of the first two harmonics together with some visible distortions in the upper part of the spectrum (the area located above 3kHz around time 0.45s corresponds to friction noise and should not be modeled by the harmonic

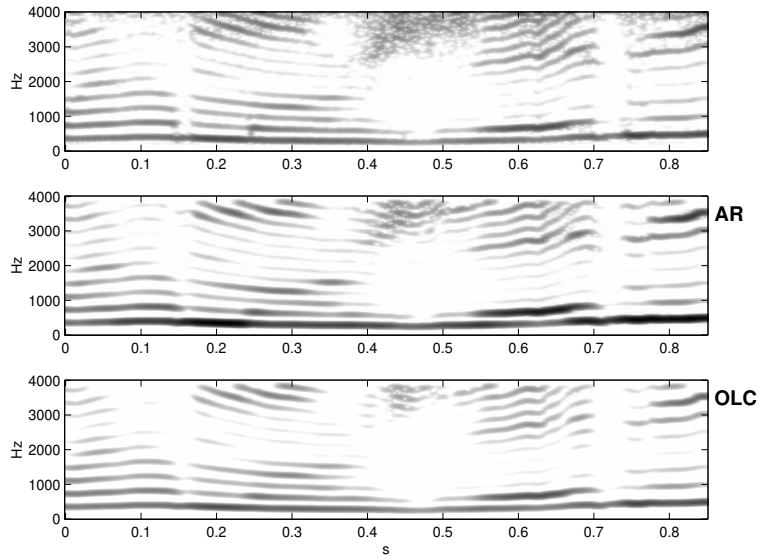

Fig. 9. Narrow band spectrograms of the original signal (top plot) and the harmonic synthetic signals obtained from AR (middle plot) and OLC (bottom plot) envelopes (pre-emphasis by $1/(1+0.97z^{-1})$, 60dB depth).

envelope). The latter problem is usually circumvented by applying the harmonic model only in the lower part of the spectrum (with a cutoff referred to as "maximum voicing frequency" in [32]), but figure 9 nonetheless shows that OLC estimation would still be useful in robustifying the harmonic envelope wrt. to errors in the determination of the maximum voicing frequency. On the example of figure 9, the distortion brought by AR envelope modeling is distinctively audible, while the synthetic time-varying harmonic signals obtained either by direct resynthesis (with the estimated harmonic magnitudes) or from the spectral envelopes estimated by the OLC or WLS methods are indistinguishable. A longer section of the signal shown in figure 9 together with the associated results and MATLAB functions needed to implement the methods discussed in section IV are available through the Internet at address `http://www.tsi.enst.fr/~cappe/env`.

V. Discussion

As already mentioned, the stationary sinusoidal model is at most an approximation of the signal behavior on a short time frame and the very concept of "actual envelope" is questionable. Among the exciting possibilities for future work, tracking of a non-stationary (or evolutive) version of the stationary sinusoidal model is certainly a key issue. Another important aspect is without doubt perception. We however feel that perceptual issues are beyond the scope of the present paper, in particular because they are application dependent: An estimation error such as the one in figure 6-A2 is largely above the perception threshold but it could be the case that in very low bit rate coding applications, this error still is concealed by the envelope distortion due to quantization or by some other source or signal

distortion. Likewise, the envelope overestimation visible on the left plot of figure 7 certainly is a problem for denoising applications where signal overestimation means increasing the residual noise level but it could be of less importance in coding applications because of masking phenomenons.

We would like however to conclude the paper by stressing that the likelihood criterion given by (11) has a strong theoretical basis (see also appendix A), provides an interesting insight into the limitation of the previously proposed method of "discrete cepstrum" (section III-B) and performs significantly better than both the AR and the discrete cepstrum methods, particularly for high pitched and noisy signals. The weighted least-squares approximation provides a reduced implementation cost alternative which is equivalent to the optimization based approach in low noise situations. On most hardware platforms, the estimator based on optimization will be suitable only for off-line applications (analysis and coding of units for speech synthesis, high quality analysis/synthesis) but the least-squares approximation can easily meet the requirement of real time speech and audio processing.

Appendix

I. Derivation of the likelihood criterion for Gaussian white noise

In this appendix, we consider the noisy harmonic model given by (1) assuming that the noise process is a Gaussian white noise with power σ^2. Under this simplifying assumption, it is shown that the likelihood associated to (1) is equivalent for large sample sizes to an expression which only depends on the observations through the Fourier estimates at the frequencies of the harmonics and an estimate of the noise power. We next show that, the likelihood integrated with respect to the phase response of the spectral envelope yields the criterion defined in (11).

The likelihood corresponding to (1) in the Gaussian case is

$$p(\mathbf{r}) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{r}-\mathbf{m})'(\mathbf{r}-\mathbf{m})\right] \tag{23}$$

where $\mathbf{r} = (r_1, \ldots, r_T)'$, $\mathbf{m} = (m_1, \ldots, m_T)'$ and the prime denote transposition. For any $T \times T$ invertible matrix $\mathbf{M}$, (23) may be rewritten as

$$p(\mathbf{r}) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{M}'(\mathbf{r}-\mathbf{m})\right)'\left(\mathbf{M}'\mathbf{M}\right)^{-1}\left(\mathbf{M}'(\mathbf{r}-\mathbf{m})\right)\right]$$

Let $\mathbf{M}$ denote the matrix defined by blocks as

$$\mathbf{M} = \left[\mathbf{P}(T \times K) \vdots \mathbf{Q}(T \times K) \vdots \mathbf{U}(T \times (T-2K))\right]$$

where $\mathbf{P}_{tk} = \cos\omega_k t$, $\mathbf{Q}_{tk} = \sin\omega_k t$ for $1 \le k \le K$, $1 \le t \le T$ and $\mathbf{U}$ is chosen among the matrices which satisfy $\mathbf{U}'\mathbf{U} = \mathbf{I}_{T-2K}$, $\mathbf{U}'\mathbf{P} = \mathbf{0}$ and $\mathbf{U}'\mathbf{Q} = \mathbf{0}$. $\mathbf{P}, \mathbf{Q}$ and $\mathbf{U}$ thus define a subspace decomposition of $\mathbb{R}'$ for which $\mathbf{U}$ is orthogonal to both $\mathbf{P}$ and $\mathbf{Q}$. For finite sample

sizes, $\mathbf{P}$ and $\mathbf{Q}$ are not orthogonal but it is a standard result that they are quasi-orthogonal for large sample sizes in the sense that [27], [31]

$$\begin{aligned}\mathbf{P}'\mathbf{P} &= \frac{T}{2}\mathbf{I}_K + O(1)\\ \mathbf{Q}'\mathbf{Q} &= \frac{T}{2}\mathbf{I}_K + O(1)\\ \mathbf{P}'\mathbf{Q} &= O(1)\end{aligned} \tag{24}$$

where the $O(1)$ notation stands for terms that can be bounded from above. Hence, for large sample sizes,

$$\begin{aligned}p(\mathbf{r}) = \frac{1}{(2\pi\sigma^2)^{T/2}}\exp\Bigg\{ \\ -\frac{T}{4\sigma^2}\left(\left[\frac{2}{T}\mathbf{P}'(\mathbf{r}-\mathbf{m})\right]'\left[\frac{2}{T}\mathbf{P}'(\mathbf{r}-\mathbf{m})\right] + \left[\frac{2}{T}\mathbf{Q}'(\mathbf{r}-\mathbf{m})\right]'\left[\frac{2}{T}\mathbf{Q}'(\mathbf{r}-\mathbf{m})\right]\right)(1+O(1/T)) \\ -\frac{1}{2\sigma^2}\left[\mathbf{U}'(\mathbf{r}-\mathbf{m})\right]'\left[\mathbf{U}'(\mathbf{r}-\mathbf{m})\right]\Bigg\}\end{aligned} \tag{25}$$

The third term in (25) normalized by the sample size T is a biased but consistent estimate of the noise power which is denoted $\hat{\sigma}^2$. Note that this term does not depend on the parameters of the envelope since the orthogonality of $\mathbf{U}$ with both $\mathbf{P}$ and $\mathbf{Q}$ implies that $\mathbf{U}'\mathbf{m} = \mathbf{0}$ so that $\hat{\sigma}^2 = \mathbf{r}'\mathbf{U}\mathbf{U}'\mathbf{r}/T$. Eq. (25) may thus be rewritten as

$$p(\mathbf{r}) = \exp\left\{-\frac{T}{4\sigma^2}\left[\left([\hat{\mathbf{a}}-\mathbf{a}]'[\hat{\mathbf{a}}-\mathbf{a}] + [\hat{\mathbf{b}}-\mathbf{b}]'[\hat{\mathbf{b}}-\mathbf{b}]\right)(1+O(1/T)) + 2\hat{\sigma}^2\right]\right\} \tag{26}$$

where $\hat{\mathbf{a}} = \frac{2}{T}\mathbf{P}'\mathbf{r}$ and $\hat{\mathbf{b}} = \frac{2}{T}\mathbf{Q}'\mathbf{r}$ respectively denote the estimated in phase and in quadrature amplitudes of the harmonics (when no data tapper is used) and $\mathbf{a} = (a_1, \ldots, a_k)'$, $\mathbf{b} = (b_1, \ldots, b_k)'$ are the actual harmonic magnitudes as defined by the envelope. In obtaining (26) we have used (24) to show that $\frac{2}{T}\mathbf{P}'\mathbf{m} \to \mathbf{a}$ and $\frac{2}{T}\mathbf{Q}'\mathbf{m} \to \mathbf{b}$. Noting that $n = \frac{4\sigma^2}{T}$ is the apparent noise level for the rectangular window, and introducing the notations of section II ($x_k = \hat{a}_k^2 + \hat{b}_k^2$, $s_k = a_k^2 + b_k^2$), (26) may be rewritten as

$$p(\mathbf{r}) = \exp\left[-\frac{T\hat{\sigma}^2}{2\sigma^2}\right]\prod_{k=1}^{K}\exp\left[\left(-\frac{s_k + x_k}{n} + 2\sqrt{\frac{x_k s_k}{n^2}}\cos(\theta_k - \hat{\theta}_k)\right)(1+O(1/T))\right] \tag{27}$$

where $\theta_k \triangleq \text{Angle}(a_k, b_k)$ denotes the phase of the kth harmonic as given by the envelope model whereas $\hat{\theta}_k \triangleq \text{Angle}(\hat{a}_k, \hat{b}_k)$ is the phase measured from the observed signal.

Note that in (27), the $O(1/T)$ terms do not depend upon any of the quantities except T itself. Assuming, that the parameters θ_k, for $1 \le k \le K$, have a prior distribution which

is uniform on $[0, 2\pi]$, it is thus possible to integrate out these nuisance parameters to obtain

$$\begin{aligned}\bar{p}(\mathbf{r}) &\triangleq \frac{1}{(2\pi)^K}\int_{[0,2\pi]^K} p(\mathbf{r})d\theta_1\dots d\theta_K \\ &= \exp\left[-\frac{T\hat{\sigma}^2}{2\sigma^2}\right]\prod_{k=1}^{K}\exp\left[-\frac{s_k+x_k}{n}\right] I_0\left(2\sqrt{\frac{x_k s_k}{n^2}}\right)(1+O(1/T)) \qquad (28)\end{aligned}$$

where I_0 is the Bessel function of order 0 defined in (8).

II. Derivatives of the likelihood criterion

In this section, we consider the first and second order derivatives of the likelihood criterion given by (11). Closed-form expressions of the gradient and Hessian are obtained that are valid for any envelope parameterization. In the case of the cepstral parameterization some arguments are provided to back up the experimental observation that the criterion is generally convex if the algorithm is started from a point sufficiently close to the true envelope parameters.

We will denote the envelope parameters by $\varphi_0, \dots, \varphi_p$ where p denotes the order of the parameterization. Differentiating (11) is made easy by the use of the following relations [1]

$$\begin{aligned}\frac{dI_0(y)}{dy} &= I_1(y) \\ \frac{d\left(yI_1(y)\right)}{dy} &= yI_0(y) \qquad (29)\end{aligned}$$

The gradient of $L(x_1, \dots, x_K|S)$ is obtained as

$$\frac{\partial L}{\partial \varphi_i} = \sum_{k=0}^{K}\left(\frac{\partial s_k}{\partial \varphi_i}\right)\frac{1}{n_k}\left[1-\sqrt{\frac{x_k}{s_k}}\frac{I_1}{I_0}\left(2\sqrt{\frac{s_k x_k}{n_k^2}}\right)\right] \qquad (30)$$

where the notation $\partial s_k/\partial\varphi_i$ is used as a short-hand for $\partial S(\omega_k)/\partial\varphi_i$. The expression of the Hessian follows:

$$\begin{aligned}\frac{\partial^2 L}{\partial\varphi_j\partial\varphi_i} &= \sum_{k=0}^{K}\left\{\left(\frac{\partial^2 s_k}{\partial\varphi_j\partial\varphi_i}\right)\frac{1}{n_k}\left[1-\sqrt{\frac{x_k}{s_k}}\frac{I_1}{I_0}\left(2\sqrt{\frac{s_k x_k}{n_k^2}}\right)\right]\right. \\ &\quad \left.-\left(\frac{\partial s_k}{\partial\varphi_j}\right)\left(\frac{\partial s_k}{\partial\varphi_i}\right)\frac{x_k}{s_k n_k^2}\left[1-\frac{n_k}{\sqrt{s_k x_k}}\frac{I_1}{I_0}\left(2\sqrt{\frac{s_k x_k}{n_k^2}}\right)-\frac{I_1^2}{I_0^2}\left(2\sqrt{\frac{s_k x_k}{n_k^2}}\right)\right]\right\} \qquad (31)\end{aligned}$$

The implementation of (30) and (31) is made easy by the fact that the only special function that needs to be evaluated is the ratio $I_1/I_0(y)$. This ratio is particularly well behaved since it is positive and for large values of y the following approximation holds [1]

$$\frac{I_1}{I_0}(y) = 1-\frac{1}{2y}+o(\frac{1}{y}) \qquad (32)$$

In general, the Hessian given by (31) is not positive definite. For the cepstral parameterization defined in (14) however, the matrix $(\partial^2 s_k/\partial c_j \partial c_i)$ is a positive rank one matrix and (31) simplifies to

$$\frac{\partial^2 L}{\partial c_j \partial c_i} = \sum_{k=0}^{K} C_{ki} C_{kj} \frac{s_k}{n_k} \left\{ 1 - \frac{x_k}{n_k} \left[1 - \frac{I_1^2}{I_0^2} \left(2\sqrt{\frac{s_k x_k}{n_k^2}} \right) \right] \right\} \tag{33}$$

where the cepstral regression matrix $\mathbf{C}$ was defined in (20). Each of the terms in the above summation has an interesting behavior when s_k (the squared amplitude of the sinusoidal component) becomes large with respect to the apparent noise level n_k: If we omit the factors that involve x_k/s_k, (32) shows that the term corresponding to the index k in (33) can be approximated as $\frac{s_k}{n_k}[\frac{1}{2}+o(1)]$. The factor x_k/s_k does not modify this result since (10) indicates that $\mathrm{E}[x_k/s_k] \to 1$ and $\mathrm{Var}[x_k/s_k] \to 0$ as $s_k \to +\infty$. Application of the continuous mapping theorem shows that the latter result is indeed valid if we use the symbol $o_p()$ which denotes convergence in probability to zero in place of $o()$ [34]. Computer simulations of this term show that it is positive with high probability even for moderate values of s_k. For instance, when the apparent signal to noise ratio s_k/n_k equals 6 dB, the estimated probability of negativeness is 0.3%.

As a consequence, if all the sinusoids are well above the apparent noise level ($s_k \gg n_k$), each of the term in (33) is non-negative with high probability, and thus the Hessian of the likelihood criterion $L(x_1, \ldots, x_K|S)$ is positive definite. Note that for the Hessian to be positive definite, it takes $K > p$ (more measurements than the number of envelope parameters) because the matrix $\mathbf{M}$ defined by $M_{ij} = C_{ki}C_{kj}$ is a rank one matrix. In practice however, the Hessian is positive definite even when this constraint isn't met, and furthermore, negative eigenvalues appear less often than suggested by the above derivations because of the constant matrix $\lambda\mathbf{R}$ (see section III-A) added by the roughness penalty which enforces the positiveness.

Because the Hessian is a continuous function of the parameters, the previous observation is true for a whole neighborhood of the actual envelope S. Thus, if the envelope is well above the noise level and if the algorithm is started from an envelope sufficiently close to the unknown true envelope, the maximization of $L(x_1, \ldots, x_K|S)$ reduces (with high probability) to a convex problem.

References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Applied Mathematics Series, 55. National Bureau of Standards, 1964.

[2] J. O. Berger, B. Liseo, and R. L. Wolpert. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28, 1999.

[3] D. M. Brookes and H. P. Loke. Modelling energy flow in the vocal tract with applications

to glottal closure and opening detection. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, volume 1, pages 213–216, 1999.

[4] M. Campedel-Oudot. *Etude du modèle "sinusoïdes et bruit" pour le traitement des signaux de parole. Estimation robuste de l'enveloppe spectrale*. PhD thesis, ENST, 1998.

[5] O. Cappé, J. Laroche, and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. In *IEEE Workshop on App. of Sig. Proc. to Audio and Acoust.*, October 1995. Paper 9a-1.

[6] O. Cappé and E. Moulines. Regularization techniques for discrete cepstrum estimation. *IEEE Signal Processing Letters*, 3(4):100–102, April 1996.

[7] J. H. Derby. Comments on "on the design of pole-zero approximations using a logarithmic error measure". *IEEE Trans. Signal Processing*, 44(7):1811–1813, July 1996.

[8] A. El-Jaroudi and J. Makhoul. Discrete all pole modeling. *IEEE Trans. Signal Processing*, 39(2):411–423, February 1991.

[9] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[10] J. L. Flanagan. *Spech analysis, synthesis and perception*. Springer-Verlag, 2nd edition, 1972.

[11] R. Di Francesco and E. Moulines. Detection of glottal closure by jumps in the statistical properties of the speech signal. In *Proc. EUROSPEECH*, pages 39–42, 1989.

[12] T. Galas and X. Rodet. An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sound signals. In *Proc. of International Computer Music Conference*, pages 82–84, Glasgow, 1990.

[13] D.W. Griffin and J.S. Lim. Multiband-excitation vocoder. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-36(2):236–243, February 1988.

[14] E. J. Hannan. The estimation of frequency. *J. Appl. Prob.*, 10:510–519, 1973.

[15] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

[16] H. Hermansky, H. Fujisaki, and Y. Sato. Analysis and synthesis of speech based on spectral transform linear predictive method. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, volume 2, pages 777–780, 1983.

[17] N. L. Johnson and S. Kotz. *Continuous Univariate Distributions*, volume 2. Wiley-Interscience, 1970.

[18] B-H. Juang, L. R. Rabiner, and J. G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, 35(7):947–954, 1987.

[19] C. H. Lee. Robust linear prediction for speech analysis. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, volume 1, pages 289–292, 1987.

[20] R. J. McAulay and T. F. Quatieri. Sinusoidal coding. In W.B. Kleijn and K.K. Paliwal, editors, *Speech Coding and Synthesis*, pages 123–176. Elsevier, 1995.

[21] R. Mizoguchi, M. Yanagida, and O. Kakusho. Speech analysis by selective linear prediction in the time domain. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, volume 3, pages 1573–1576, 1982.

[22] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518, 1986.

[23] M. Oudot, O. Cappé, and E. Moulines. Robust estimation of the spectral envelope for "harmonics + noise " models. In *IEEE Workshop on speech coding*, Pocono Manor, September 1997.

[24] D.B. Paul. The spectral envelope estimation vocoder. *IEEE Trans. Acoust., Speech, Signal Processing*, 29(4):786–7941, august 1981.

[25] Y. Pawitan and F. O'Sullivan. Non parametric spectral density estimation using penalized whittle likelihood. *Journal of the American Statistical Association*, 89(426):600–610, june 1994.

[26] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical recipes in C : the art of scientific computing*. Cambridge University Press, second edition, 1992.

[27] B. G. Quinn and P. J. Thomson. Estimating the frequency of a periodic function. *Biometrika*, 78(1):65–74, 1991.

[28] L. R. Rabiner and B-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[29] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals*. Prentice-Hall, 1978.

[30] X. Serra. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. PhD thesis, CCRMA Department of Music, Stanford University, Stanford, California, 1989. Report No. STAN-M-58.

[31] P. Stoica and R. Moses. *Introduction to spectral analysis*. Prentice Hall, 1997.

[32] Y. Stylianou, J. Laroche, and E. Moulines. High-quality speech modification based on a harmonic + noise model. In *Proc. EUROSPEECH*, pages 451–454, Madrid, September 1995.

[33] Y. Tohkura, F. Ikatura, and S. Hashimoto. Spectral smoothing techniques in PARCOR speech analysis-synthesis. *IEEE Trans. Acoust., Speech, Signal Processing*, 26(6):587–596, 1978.

[34] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1988.

[35] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, 68(5):1523–1525, 1980.