Regularized Estimation of Cepstrum Envelope from Discrete Frequency Points

O. Cappé, J. Laroche, E. Moulines

Ecole Nationale Supérieure des Télécommunications
Déepartement Signal / CNRS-URA 820
46 Rue Barrault 75634 Paris Cedex 13, FRANCE.
email: laroche@sig.enst.fr

ABSTRACT

This paper presents an improved method for the estimation of a continuous frequency-envelope when the value of this envelope is specified only at discrete frequencies. It is based on the Galas/Rodet approach which consists of fitting a cepstral amplitude envelope to the specified frequency points by minimizing a frequency-domain least-squares criterion. This paper introduces a regularization technique which increases the robustness of the estimation procedure. Used in combination with a warped frequency-scale, the proposed method is shown to provide an efficient model for the frequency envelope of speech signals.

1. Introduction

This paper deals with the problem of estimating a continuous frequency-envelope when the value of this envelope is specified only at discrete frequencies. This problem arises naturally in sinusoidal analysis/synthesis systems in which the signal is modeled as the discrete sum of sinusoids. Such a continuous envelope is needed for example for pitch-scale modifications of speech signals (because the amplitudes of the modified harmonics must be extrapolated from the knowledge of the original harmonic amplitudes).

A number of methods for amplitude envelope estimation have been proposed. El-Jaroudi and Makhoul suggested to fit an all-pole transfer function to the discrete set of frequency points, but the minimization stage requires the use of a costly iterative procedure [4]. McAulay and Quatieri's technique consists of linearly interpolating the discrete frequency points, then applying a standard cepstrum modeling method to this continuous interpolated envelope [9]. However, the method yields accurate envelopes only when a sufficient number of cepstral coefficients are used. Galas and Rodet proposed to estimate directly the cepstral coefficients through the minimization of a frequency-domain least-squares criterion [5]. As will be shown below, the method proves very efficient but is plagued with ill-conditioning problems.

The advantages of the cepstral coefficients representation [10] are numerous: it was found to provide a perceptually-realistic distance measure for assessing the similarity of sound envelopes, making it a natural candicate for speech/speaker recognition problems; it usually provides smooth envelopes (by contrast with autoregressive envelope modeling), which is a desirable feature in the context of speech synthesis.

The method proposed in this contribution is based on the Galas/Rodet approach, but makes use of a regularization technique to increase the robustness of the estimation procedure, providing for the use of warped frequency scales.

2. Basic discrete cepstrum

2.1. Description

In the following, we will assume that the amplitude envelope is known at L discrete normalized frequencies f_k. We will denote a_k the amplitude envelope measured at frequency f_k. A typical case is the analysis of quasi-periodic signals such as speech or music, in which the frequencies f_k are the harmonics of the fundamental frequency f_1 and the harmonic amplitudes are determined by a Fourier analysis [1] or a least-squares minimization [2, 3].

The log-amplitude envelope $A_c(f)$ in dB will be described by the so called 'real cepstrum parameters' c_i and take the form

$$A_c(f) = c_0 + 2\sum_{i=1}^{p} c_i \cos(2\pi f i) \qquad (1)$$

As is well known, this expression is consistent with the definition of the cepstrum as the inverse Fourier transform of the logarithm of the modulus of the signal's Fourier transform. In the following we will consider that the order p of the cepstrum is known a priori.

The problem is now to determine the set of cepstrum coefficients c_i such that the log-amplitude envelope $A_c(f)$ in dB evaluated at frequencies f_k is maximally close to the desired amplitudes a_k. This can be expressed by a weighted least-squares criterion which takes the simple form

$$\epsilon = \sum_{k=1}^{L} w_k ||20\log_{10} a_k - A_c(f_k)||^2, \qquad (2)$$

in which w_k are weights that can be used to obtain a better fit at certain discrete frequencies [4]. This criterion can be expressed in a matrix form as

$$\epsilon = ||a - Mc||_W^2 = (a - Mc)^T W(a - Mc) \qquad (3)$$

where

$$a = 20\left[\log_{10}(a_1)\ldots\log_{10}(a_L)\right]^T$$

$$M = \begin{bmatrix} 1 & 2\cos(2\pi f_1) & 2\cos(2\pi f_1 2) & \ldots & 2\cos(2\pi f_1 p) \\ \vdots & \vdots & & \vdots & \\ 1 & 2\cos(2\pi f_L) & 2\cos(2\pi f_L 2) & \ldots & 2\cos(2\pi f_L p) \end{bmatrix}$$

and **c** is the vector of unknown cepstrum parameters:

$$c = [c_0 \ldots c_p]^T$$

and W is a diagonal matrix with diagonal elements $[w_1, \ldots, w_L]$. The least-squares solution is easily found to be

$$c = (M^T W M)^{-1} M^T W a \tag{4}$$

i.e., the cepstrum coefficients are obtained by a simple matrix inversion (provided $p < L$).

2.2. Problems associated with the standard method

When used as described above, the standard discrete cepstrum method is known to yield meaningless results because the matrix $M^T W M$ is frequently poorly conditioned [7]. This means that non-significant perturbations of the data (f_k or a_k) such as machine rounding errors can induce very large variations of the estimated cepstrum coefficients and of the log-amplitude envelope $A_c(f)$. Fig. 1 presents such a case: the frequency points (the circles in the figure) correspond to the amplitudes of the harmonics of a voiced speech signal. The resulting amplitude envelope obviously lacks smoothness and even takes abnormally high and low values in the high frequency range. It is also observed that the envelope tends to have important oscillations in the region where the amplitude differences between successive points are large. In this case, the condition number of matrix $M^T W M$ was found to be of the order of 10^5. In practice, such problems occur frequently, especially 1) when there are broad frequency regions in which no frequency point is specified, 2) when some frequency points are closely spaced in frequency but very different in magnitude, or 3) when the number cepstral coefficients approaches the number of frequency points. In addition, the standard method requires the number of frequency points to be larger than the number of cepstral coefficients (otherwise, matrix $M^T W M$ is singular), which can be a problem in practice as the the number of frequency points is usually related to the pitch of the signal.

3. Regularized discrete cepstrum

To overcome the problems associated with the standard discrete cepstrum method, Galas and Rodet [5] suggested to increase the number L of frequency points by replacing the original f_k by clusters of neighboring points. In many cases, this technique yields satisfying results, at the price of an increased

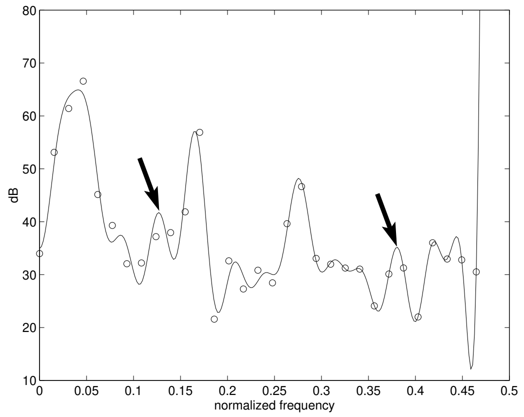

Figure 1: Log-amplitude envelope estimated by the direct minimization of the least-squares criterion (with an identity weighting matrix W) for a cepstrum of order p = 27. The circles show the specified frequency points.

numerical complexity. However, this technique breaks down in the case 1) described in the preceding section. Moreover, the resulting envelope depends significantly on the choice and number of points in the clusters, as well as other parameters (weights etc...).

By contrast, the method proposed in this this paper is based on a well-known regularization technique which consists of imposing additional constraints on the log-amplitude envelope. The idea consists in seeking an envelope which, in addition to minimizing the least-squares criterion Eq. (2), is also *smooth*, in a sense that will be precized below. Following [8], the least-squares criterion is modified as follows:

$$\epsilon_r = \sum_{k=1}^{L} w_k ||20\log_{10} a_k - A_c(f_k)||^2 + \lambda \mathcal{R}[A_c(f)] \tag{5}$$

where $\mathcal{R}[A_c(f)]$ is a penalty functional: $\mathcal{R}$ is small if the envelope is smooth, and large in the other case. λ is the regularization parameter which controls the relative importance of the smoothness constraint in the criterion to be minimized. As indicated by Eq. (5), the new criterion favors envelopes that are close to the specified frequency points (first term in the right member of Eq. (5)) while exhibiting some degree of smoothness (second term in the right member of Eq. (5)).
A possible smoothness criterion is

$$\mathcal{R}[A_c(f)] = \int_{-1/2}^{1/2} \left[\frac{d}{df} A_c(f)\right]^2 df \tag{6}$$

which is null when $A_c(f)$ is a constant. Fortunately, this smoothness criterion can be expressed as a quadratic form of the cepstral coefficients: inserting Eq. (1) in Eq. (6) and using

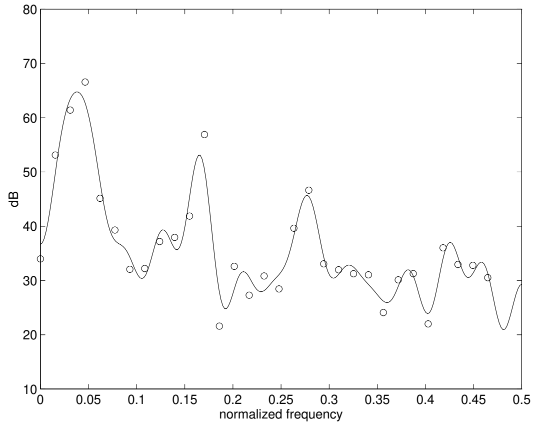

Figure 2: Log-amplitude envelope estimated by the regularized least-squares procedure (with an identity weighting matrix) for a cepstrum of order p = 27, with $\lambda = 5e - 4$. The circles show the specified frequency points. Arrows indicate over-oscillations.

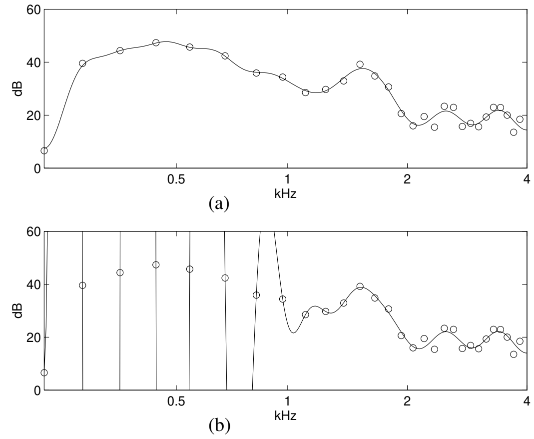

Figure 3: Estimated log-amplitude envelope with a Bark frequency scale; **(a)** with the regularized least-squares procedure ($\lambda = 5e - 4$); **(b)** with the direct least-squares minimization. The cepstrum order p is set to 20 and an identity weighting matrix is used in both cases.

straightforward manipulations, one finds that

$$\mathcal{R}[A_c(f)] = c^T Rc \tag{7}$$

where R is a diagonal matrix whose diagonal elements are $8\pi^2[0, 1^2, 2^2, \ldots, p^2]$. Finally, the solution to the modified criterion Eq. (5) is given by

$$c = \left(M^T WM + \lambda R\right)^{-1} M^T Wa \tag{8}$$

Fig. 2 displays the results of the regularized technique for the same signal as in Fig. 1. The problem visible in the high frequency range in the standard method has disappeared. Closer inspection reveals that unnecessary oscillations (indicated by the arrows in Fig. 1) have also been removed, and that the envelope fit at some of the specified frequency points is slightly worse than in the standard method. In this case, the condition number of matrix $\left(M^T WM + \lambda R\right)$ was found to be about 60. In practice, values of the regularization parameter λ of the order of 10^{-4} make it possible to eliminate problems associated with ill-conditioning, while maintaining a good envelope fit.

4. Warped frequency scale

It is common practice, in many cases, to use a warped (rather than linear) frequency axis. Examples include speech coding [9], speech recognition [10] or filter design [11]. The idea consists of enlarging the low-frequency region relative to the upper frequency range. Standard warped scales include the logarithmic scale and the Bark scale [12]. In the following, we will denote by $f \to f' = G(f)$ the frequency warping function. The normalized warped frequency lies in the range $0 < f' < 0.5$.

Frequency warping in the discrete cepstrum method has been proposed by Galas and Rodet [6] under the name 'discrete MFCC'. The modification of the algorithm is quite straightforward: In the expression of matrix M, the frequencies f_k are simply replaced by $G(f_k)$. Unfortunately, when frequency warping is used in the standard discrete cepstrum method, ill-conditioning almost systematically occurs, as exemplified by Fig. 3, bottom curve, in the case of a Bark frequency scale. By contrast, the regularized method performs as well with warped or linear frequency scales (Fig. 3, top curve).

For speech processing, it was found in practice that using a warped frequency scale to improve the fit in the low-frequency range is by far more efficient than using a weighting matrix W. This is especially true when the frequency points are close to each other, for example in the case of low-pitched signals, because rapid variations of the log-amplitude envelope cannot be obtained when few cepstral coefficients are used, whatever the weighting function. Rather than increasing the number of cepstral coefficients, the frequency points can be spread further apart in a given frequency range, thus facilitating the fit in that area. This is clearly seen in the upper part of Fig. 3: the fit is extremely good in the lower frequency range, but becomes looser as frequency increases and frequency points get closer to one another.

Fig. 4 presents the comparison of results obtained by the regularized discrete cepstrum method for a linear and a Bark frequency scale. Approximately 20s of voiced speech signal were processed, extracted from the voice of a male speaker with a low fundamental frequency (between 95Hz and 100Hz). The pitch was found in a first step, then used to

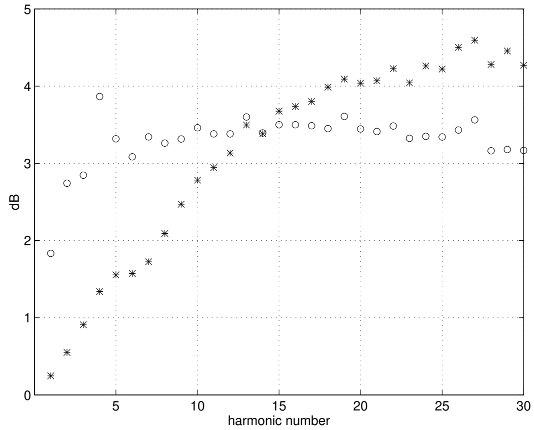

Figure 4: Standard deviation of the modeling error (in dB) as a function of the harmonic number; circles correspond to the use of a linear frequency scale; stars correspond to the use of a Bark frequency scale.

estimate the harmonic amplitudes every 10ms, yielding 2000 sets of about 40 frequency points each (the speech was sampled at 8kHz). Each set of frequency points was modeled by use of the regularized discrete cepstrum method, using 17 cepstral coefficients, for a linear then a Bark frequency scale. For each harmonic, the modeling error was calculated as

$$E_k = 20 \log_{10} a_k - A_c(f_k)$$

and its standard deviation across all 2000 analyses is plotted in Fig. 4 as a function of the harmonic rank k. No weighting was used. When the linear frequency scale is used, (circles in Fig. 4) the standard deviation of the modeling error is nearly the same for all harmonics, lying between 3 and 4 dB. As expected, when the Bark frequency scale is used, the fit is better in the low-frequency area and slightly worse in the high-frequency area. In both cases, even though the cepstrum order is relatively small (16), the modeling error is reasonably low.

To check the accuracy of the discrete spectrum technique, the modeled amplitudes $A_c(f_k)$ were used in a sinusoidal synthesis system (the HNM synthesis system [3]) in place of the original amplitudes $20 \log_{10} a_k$. The experiment demonstrated that a very good low-frequency fit is necessary to preserve the voice presence. The voice synthesized using the amplitudes obtained by the linear frequency-scale analysis was still of very good quality but slightly differed from the original voice in its ‘presence’ quality. The voice synthesized using the amplitudes obtained by the Bark frequency-scale analysis was indistinguishable from the original voice.

5. Conclusion

The Bark frequency-scale, regularized cepstrum method has been used successfully for speech analysis/synthesis in the context of the HNM (Harmonic plus Noise Model) system. It was found to be much more robust for pitch-scale modifications than either autoregressive, piecewise linear, or the standard discrete cepstrum envelope modeling. This method is also currently used in a voice-conversion system (a system used to transform the voice of a given speaker into that of another one) [13]. The method could also prove useful in the context of speech/speaker recognition, as well as for coding purposes.

References

1. R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, Aug 1986.
2. L.B. Almeida and F.M. Silva. Variable-frequency synthesis: an improved harmonic coding scheme. *Proc. IEEE ICASSP-84*, pages 27.5.1–27.5.4, 1984.
3. J. Laroche, E. Moulines, and Y. Stylianou. HNS: Speech modification based on a harmonic + noise model. *Proc. IEEE ICASSP-93, Minneapolis*, Apr 1993.
4. A. El-Jaroudi and J. Makhoul. Discrete all pole modeling. *IEEE Trans. Acoust., Speech, Signal Processing*, 39(2):411–423, Fev 1991.
5. T. Galas and X. Rodet. An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sounds. *Proc. of International Computer Music Conference, Glasgow*, pages 82–84, 1990.
6. T. Galas and X. Rodet. Generalized functional approximation for source-filter system modeling. *Proc. Eurospeech, Genova*, pages 1085–1088, 1991.
7. C. L. Lawson and R. J. Hanson. *Solving Least–Squares Problems*. Prentice Hall, Englewood Cliffs, New Jersey, 1974.
8. A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. Scripta series in mathematics. Winston, Washington, 1977.
9. R. J. McAulay and T. F. Quatieri. Low-rate speech coding based on the sinusoidal model. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, chapter 6, pages 165–208. Marcel Dekker, 1991.
10. L. R. Rabiner and B-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.
11. J. O. Smith. *Techniques for Digital Filter Design and System Identification with Application to the Violin*. PhD thesis, Stanford University, Stanford, CA, Jun 1983.
12. 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.
13. Y. Stylianou, O. Cappé, and E. Moulines. Statistical methods for voice quality transformation. *Accepted, EUROSPEECH 95*, September 1995.