HIGH RESOLUTION SPHERICAL QUANTIZATION OF SINUSOIDAL PARAMETERS USING A PERCEPTUAL DISTORTION MEASURE

Pim Korten, Jesper Jensen and Richard Heusdens

Information and Communication Theory Group, Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
phone:+31 (0)15 27 82188, fax:+31 (0)15 27 81843
email: {p.e.l.korten, j.jensen, r.heusdens}@ewi.tudelft.nl

ABSTRACT

Sinusoidal modelling is a key technology in low rate audio coding, and methods for efficient quantization of sinusoidal parameters are therefore of high importance. In this work we derive analytical formulas for the optimal entropy constrained unrestricted spherical quantizers for amplitude, phase and frequency, using a perceptual distortion measure. This is done both for a single sinusoid, and for multiple sinusoids distributed over multiple segments. The quantizers minimize a high-resolution approximation of the expected distortion, while the corresponding quantization indices satisfy an entropy constraint. The quantizers turn out to be flexible and of low complexity, in the sense that they can be determined easily for varying bit rate requirements, without any sort of retraining or iterative procedures. In objective and subjective comparison tests, the proposed method is shown to outperform an existing state-of-the-art sinusoidal quantization scheme, where quantization of frequency parameters is done independently.

1. INTRODUCTION

Parametric coding is an efficient tool for representing audio signals at low bit rates [1, 2]. A common parametric model decomposes an audio signal into three components: a sinusoidal component, a noise component and a transient component, each of which are coded by different subcoders. The sinusoidal component, represented by amplitude, phase and frequency parameters, is perceptually the most important of the three. Consequently, in most low-rate audio coders the main part of the bit budget will be assigned to this component [2]. The bit budget available for encoding of sinusoids is typically given a priori e.g. by a rate-distortion control algorithm which distributes the total bit rate over the subcoders. Therefore, it is desirable to develop simple and flexible quantizers which can adapt to changing bit rate requirements without any sort of retraining or iterative procedures. In this work we focus on deriving such efficient quantizers for the sinusoidal component, and its corresponding parameters.

The quantization scheme that is presented in this work is called entropy constrained unrestricted spherical quantization (ECUSQ), and is an extension of ECUPQ (entropy constrained unrestricted polar quantization), introduced in [3]. While in ECUPQ only amplitude and phase quantization was considered, in ECUSQ all three sinusoidal parameters are quantized. The term unrestricted refers to the fact the sinusoidal parameters are quantized dependently. In [3, 4], optimal quantizers are derived under a high-resolution assumption, i.e. a very large number of quantization cells, which implies that the probability density functions of the input variables can assumed to be constant in each quantization cell. Although the derived quantizers in [4] are flexible and of low complexity, the distortion measure used is an ℓ_2 measure, so perceptual effects are not taken into account.

In this work, optimal scalar quantizers are derived for ECUSQ, using a perceptual distortion measure. More specifically, under high-resolution assumptions, optimal amplitude, phase and frequency quantizers are derived which minimize the expected distortion, while satisfying an entropy constraint. This is done both for a single sinusoid, and for the more practically relevant setting with multiple sinusoids distributed across multiple segments. For this last case, a comparison is made between ECUSQ and ECUPQ. The main advantage of the proposed method over ECUPQ is that the bit distribution between amplitude, phase and frequency does not need to be given a priori, but follows as a result of the derived formulas. In contrast, ECUPQ describes the optimal bit distribution between amplitude and phase, but does not specify the share of the bit budget that should be assigned to frequency, this needs to be chosen a priori.

2. ENTROPY CONSTRAINED UNRESTRICTED SPHERICAL QUANTIZATION

2.1. High-rate expression for the average distortion - single sinusoid

In this section we derive a high-resolution approximation for the average distortion in the case where the target signal consists of one single sinusoid, using the perceptual distortion measure. Later, we will generalize our results and consider the more practically relevant situation where the target signal is represented by several sinusoids. The perceptual distortion measure is given by

$$d(\varepsilon) = \int_0^1 \hat{a}(f)|\hat{w\varepsilon}(f)|^2 df, \tag{1}$$

and was introduced in [5]. In (1) $\varepsilon(n)$ denotes the difference between the original and quantized spherical representation of a complex sinusoid, denoted by $ae^{j(\nu n+\phi)}$ and $\tilde{a}e^{j(\tilde{\nu} n+\tilde{\phi})}$ respectively, for $n = n_0, \ldots, n_0+N-1$. Here a, ϕ and ν are amplitude, phase and frequency respectively, $n_0 \in \mathbb{Z}$, and N is the framelength. Furthermore, w is the analysis window used and $\hat{a}(f)$ is defined

The research is supported by STW, applied science division of NWO and the technology programme of the Dutch ministry of Economic Affairs.

as the inverse of the masking threshold sampled at frequency f. For $N \to \infty$ the power spectrum of the windowed error signal $|\hat{w\varepsilon}(f)|^2$ will converge to a sum of delta-functions at frequencies ν and $\tilde{\nu}$. This, in turn, means that the side lobes of $|\hat{w\varepsilon}(f)|^2$ can be neglected, and the widths of the main lobes are sufficiently small to assume the masking threshold to be constant across a main lobe. Due to high-resolution assumptions and the slowly changing masking curve, we can also assume that $\hat{a}(\nu) \approx \hat{a}(\tilde{\nu})$. Hence we can approximate $d(\varepsilon)$ by

$$d(\varepsilon) \approx \hat{a}(\tilde{\nu}) \int_0^1 |\hat{w\varepsilon}(f)|^2 df = \hat{a}(\tilde{\nu}) \sum_{n=n_0}^{n_0+N-1} |w(n)\varepsilon(n)|^2 ,$$

for N sufficiently large. With this approximation, the expected perceptual distortion is given by

$$D = E\left(d(\varepsilon)\right) = \iiint f_{A,\Phi,F}(a,\phi,\nu) d(\varepsilon) da d\phi d\nu, \quad (2)$$

where $f_{A,\Phi,F}(a,\phi,\nu)$ is the joint probability density function of amplitude, phase and frequency. In the same way as done in [4], we obtain the following high-resolution approximation for the expected distortion

$$\begin{aligned} D \approx & \frac{\|w\|^2}{12} \iiint f_{A,\Phi,F}(a,\phi,\nu)\hat{a}(\nu)\Big(g_A^{-2}(a,\phi,\nu) \\ & + a^2\left(g_\Phi^{-2}(a,\phi,\nu) + \sigma^2 g_F^{-2}(a,\phi,\nu)\right)\Big) d\nu d\phi da, \end{aligned} \quad (3)$$

where $\sigma^2 = \frac{1}{\|w\|^2}\sum_{n=n_0}^{n_0+N-1} w(n)^2 n^2$. In this derivation we used high-resolution assumptions and hence substituted quantization step sizes by so-called quantization point density functions [6, 7] g_A, g_Φ and g_F, which when integrated over a region S gives the total number of quantization levels within S. In the case of scalar quantizers, this means that the quantizer step sizes are just given by the reciprocal values of the point densities, that is, $g = \Delta^{-1}$. In high-resolution theory, quantizers are described by these density functions, without exactly specifying the location of the quantization points. Note that since we consider unrestricted quantization, the quantization point density functions depend on all three parameters.

2.2. Entropy-constrained minimization of the average distortion - single sinusoid

In this section we determine the quantization point density functions that solve

$$\min_{g_A, g_\Phi, g_F} D \text{ subject to } H(\tilde{A},\tilde{\Phi},\tilde{F}) \le H_t \quad (4)$$

where H_t is the given total target entropy, and $H(\tilde{A},\tilde{\Phi},\tilde{F})$ is the joint entropy of amplitude, phase and frequency quantization indices where $\tilde{A}$, $\tilde{\Phi}$ and $\tilde{F}$ denote their corresponding alphabets, respectively. As in [4] the joint entropy $H(\tilde{A},\tilde{\Phi},\tilde{F})$ can be approximated, under high-resolution assumptions, by

$$\begin{aligned} H(\tilde{A},\tilde{\Phi},\tilde{F}) \approx & \, h(A,\Phi,F) \\ & + \iiint f_{A,\Phi,F}(a,\phi,\nu)\log(g_A(a,\phi,\nu)) d\nu d\phi da \\ & + \iiint f_{A,\Phi,F}(a,\phi,\nu)\log(g_\Phi(a,\phi,\nu)) d\nu d\phi da \\ & + \iiint f_{A,\Phi,F}(a,\phi,\nu)\log(g_F(a,\phi,\nu)) d\nu d\phi da, \end{aligned} \quad (5)$$

where $h(A,\Phi,F)$ is the joint differential entropy of amplitude, phase and frequency, which is independent of the quantization point density functions. The constrained minimization problem in (4) can be turned into an unconstrained problem, using the method of Lagrange multipliers. In this way we obtain a Lagrangian cost function $J = D + \lambda H$, where λ is the Lagrangian multiplier. Minimizing this cost function by evaluating the Euler-Lagrange equations for all three quantization point densities, we obtain the optimal high-resolution ECUSQ quantizers for the perceptual distortion measure:

$$g_A(a,\phi,\nu) = g_A(\nu) = C(\nu) 2^{\frac{1}{3}\left(\tilde{H} - 2b(A) - \log(\sigma)\right)}, \quad (6)$$

$$g_\Phi(a,\phi,\nu) = g_\Phi(a,\nu) = a g_A(\nu), \quad (7)$$

$$g_F(a,\phi,\nu) = g_F(a,\nu) = \sigma a g_A(\nu), \quad (8)$$

where $\tilde{H}_t = H_t - h(A,\Phi,F)$ and $b(A) = \int f_A(a)\log(a) da$ are introduced for notational convenience. Furthermore

$$C^2(\nu) = \hat{a}(\nu) 2^{-d(F)}, \quad (9)$$

where $d(F) = \int f_F(\nu)\log(\hat{a}(\nu)) d\nu$. Note that since $C^2(\nu)$ is proportional to $\hat{a}(\nu)$, the perceptually more important sinusoids are quantized more finely, as compared to the ℓ_2-case [4] where quantization is frequency independent. The minimal expected distortion for ECUSQ in the perceptual case can now be found by substituting (6), (7) and (8) in (3):

$$D_{ECUSQ} = \frac{\|w\|^2 2^{-\frac{2}{3}\left(\tilde{H}_t - 2b(A) - \frac{3}{2}d(F) - \log(\sigma)\right)}}{4}. \quad (10)$$

It is easy to verify that all three parameters give exactly the same contribution to this distortion. Furthermore, it is not difficult to show that if w is an evenly symmetric window, the distortion (10) is minimal for $n_0 = -\frac{(N-1)}{2}$. We assume this to be the case in the remainder of this work.

2.3. Multiple sinusoids in multiple segments

In sinusoidal coding an input signal is split up into a number of consecutive segments of variable length and each segment is then modelled as a sum of sinusoids. The quantization distortion in a segment consists of the quantization distortion of the individual components plus a contribution due to the mutual interaction of the components. As shown in [8] this mutual interaction can be neglected if the sinusoids are spaced sufficiently far apart in the frequency domain. For practical purposes this is the case if the sinusoids are estimated using the psycho-acoustical matching pursuit algorithm [9]. We also assume that entropy and distortion are additive and independent over segments. Let K denote the number of segments, with L_k denoting the number of sinusoidal components in segment k. For the total quantization distortion we then have

$$D \approx \sum_{k=1}^{K}\sum_{l=1}^{L_k} D_{k,l} \quad (11)$$

where

$$\begin{aligned} D_{k,l} = & \frac{\|w_k\|^2}{12} \iiint f_{A,\Phi,F}(a,\phi,\nu)\hat{a}_k(\nu)\Big(g_{A_{k,l}}^{-2}(a,\phi,\nu) \\ & + a^2\left(g_{\Phi_{k,l}}^{-2}(a,\phi,\nu) + \sigma_k^2 g_{F_{k,l}}^{-2}(a,\phi,\nu)\right)\Big) d\nu d\phi da, \end{aligned}$$

and $\sigma_k^2 = \frac{1}{\|w_k\|^2} \sum_{n=-\frac{N_k-1}{2}}^{\frac{N_k-1}{2}} w_k(n)^2 n^2$, with N_k the length of segment k. Furthermore w_k is the analysis window used in segment k, and $\hat{a}_k(\nu)$ is the inverse of the masking curve corresponding to segment k. We assume the joint probability density $f_{A,\Phi,F}(a,\phi,\nu)$ to be the same for all segments.

Summing (5) for each sinusoid in each segment, we can obtain an expression for the total entropy in the same way. As in the previous section we wish to determine the quantizers which minimize the total distortion (11), such that the resulting total entropy is at a pre-specified target entropy R_t. Thus we have a constrained minimization problem which can be solved in the same way as in the previous section, giving us the following expressions for the optimal quantizers in this case:

$$\begin{aligned} g_{A_{k,l}}(a,\phi,\nu) &= g_{A_k}(\nu) \\ &= \|w_k\| C_k(\nu) 2^{-\frac{1}{3}(h(A,\Phi,F)+2b(A))} \\ &\quad \times 2^{\frac{1}{3}\gamma^{-1}\left(R_t - \sum_{m=1}^{K} L_m(\log(\sigma_m)+3\log(\|w_m\|))\right)}, \\ g_{\Phi_{k,l}}(a,\phi,\nu) &= g_{\Phi_k}(a,\nu) = a g_{A_k}(\nu), \\ g_{F_{k,l}}(a,\phi,\nu) &= g_{F_k}(a,\nu) = \sigma_k a g_{A_k}(\nu), \end{aligned}$$

where

$$C_k(\nu)^2 = \hat{a}_k(\nu) 2^{-\gamma^{-1}\sum_{m=1}^{K} L_m d(F_m)},$$

and $\gamma = \sum_{k=1}^{K} L_k$ and $d(F_k) = \int f_F(\nu) \log(\hat{a}_k(\nu)) d\nu$.

3. EXPERIMENTAL RESULTS

3.1. Validity of the high-resolution approximation of the expected distortion

In this section the theoretical rate-distortion approximation derived in (10) for a single sinusoid, is compared to a rate-distortion curve, which is practically obtained by generating a large number of realizations of single sinusoids, quantizing these sinusoids with the derived quantizers for different target bit rates H, and measuring the resulting quantization distortion. To do this, first let X, Y and Z denote three independent Gaussian variables, with zero mean and a variance of 1000.[1] Transforming these variables to the spherical domain, and using the rules for computing probability density functions of a transformation of random variables, it can be shown that the amplitude A is Maxwell distributed, the phase Φ is uniformly distributed on $[0, 2\pi]$, and the frequency F has a probability density function given by $f_F(\nu) = \frac{sin(\nu)}{2}$ for $0 \leq \nu \leq \pi$. A large number of triplets $\{a, \phi, \nu\}$ is generated from these distributions, and for each triplet the corresponding masking threshold $\hat{a}(f)$ and then the value $C(\nu)$, given in (9), is computed, using a Hanning window of length 1024. Subsequently, the triplets are quantized using (6), (7) and (8) for a given target entropy. The quantization distortion (1) for each triplet is determined, and averaged over all triplets. Repeating this procedure for several different target entropies H_t, we obtain a practical rate distortion curve which is plotted in Figure 1, where we used 10000 triplets. In the same figure the theoretical rate distortion curve given by (10) is plotted. Clearly, the curves converge towards each other, which verifies that the expression (10) for the average distortion is indeed a good approximation at high rates. For low rates it is clear that the approximation does not hold anymore.

[1] This variance is close to the one experienced in the simulation experiments with real audio data, described in the following subsection

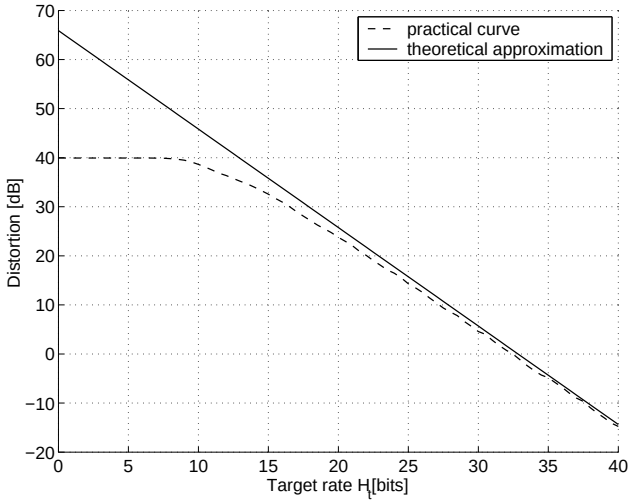

Fig. 1. Theoretical versus practical distortion-rate performance for $N = 1024$.

3.2. Comparison with ECUPQ

In this section we compare the proposed scheme with ECUPQ [3], for multiple sinusoids distributed over multiple segments, using the perceptual measure. To be able to make this comparison, we first derive the optimal ECUPQ amplitude and phase quantizers, for a certain target entropy R_{t_1}. Secondly, consider the problem of independent quantization of frequency, and derive the corresponding optimal entropy constrained frequency quantizer at a target entropy R_{t_2}, where $R_{t_1} + R_{t_2} = R_t$, the total target entropy in the ECUSQ scheme. In this way, we obtain a second scheme of three quantizers, where amplitude and phase are quantized dependently, but independently of frequency. The most important advantage the proposed method offers over the ECUPQ method is that in the proposed scheme the bit distribution between amplitude, phase and frequency follows directly from the derived formulas. In contrast, as ECUPQ was derived for optimal quantization of amplitude and phase, ECUPQ describes the optimal bit distribution between these two parameters, but does not specify the share of the total bit budget that should be assigned to frequency parameters.

Given a real audio signal, we determine the optimal segmentation and the optimal number of sinusoids on each segment, such that the modelling distortion is minimal. Here the sinusoids are estimated using the psycho-acoustical matching pursuit algorithm [9], and the optimal segmentation and distribution of sinusoids is formed using the dynamic programming based algorithm in [10]. The resulting segmentation and distribution of sinusoids, is the fixed input to our quantizer schemes. After quantizing the sinusoids with both schemes, we can measure the total distortion by computing the perceptual error (1) for each sinusoid, and then adding these errors over segments and sinusoids, where we use the assumption that distortions are additive. In Figure 2 the total distortions of the two schemes are plotted against the target rate per sinusoidal component, for a female speech signal sampled at 44.1 kHz. In this plot, five different bit distributions are considered for the ECUPQ scheme, where distributions are represented by the percentage of the rate that is assigned to frequency. Clearly, the proposed scheme performs better for this audio fragment, independent of what distribution is chosen in the ECUPQ scheme. Note also that the optimal distribution in the ECUPQ scheme is

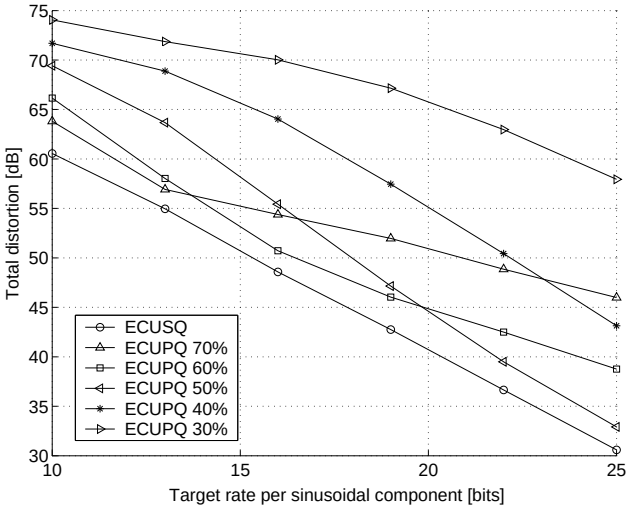

Fig. 2. ECUSQ versus ECUPQ for a female speech signal sampled at 44.1 kHz.

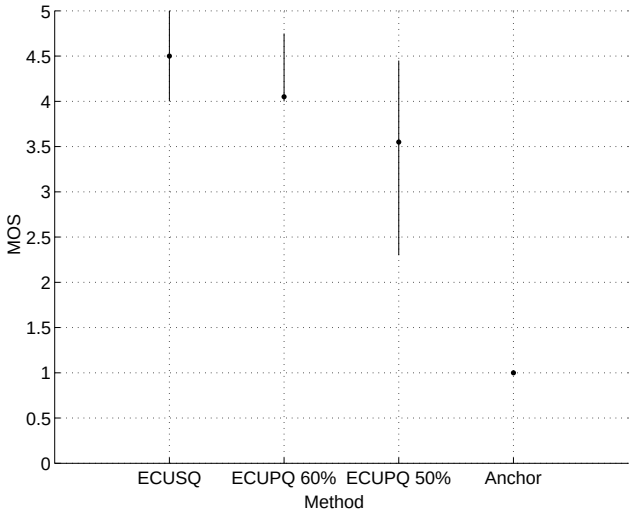

Fig. 3. Results of the listening test at 14 bits per sinusoid.

dependent on the bit rate, i.e. no fixed distribution is optimal. This is a considerable problem since the distribution needs to be chosen a priori in this method. Furthermore, even if the optimal bit distribution for ECUPQ is known (i.e. by doing an exhaustive search), the proposed scheme still performs better, which is due to jointly optimizing for all three parameters.

Since we use a perceptual distortion measure, a listening test was performed to compare both schemes. Six excerpts were included in this test: castanets, contemporary pop, Celine Dion, harpsichord, Carl Orff (classical) and female speech, all sampled at 44.1 kHz. For each excerpt, the optimal segmentation and distribution of sinusoids was determined, resulting in a certain modelling distortion. After quantizing the modelled signals with both schemes, the modelling error was added, such that the difference between the resulting signals and their original versions is due to quantization distortion. In the listening experiment, the original signal was put in as a reference (known to the listeners), in comparison to which the participants ranked the quantized versions of the signal from 1 (very poor) to 5 (no difference with the original). The target rate per sinusoid was set at 14 bits. For each fragment 3 different quantized versions were included in the test: ECUSQ, ECUPQ at a distribution of 60% (approximately the optimal distribution at 14 bits), and ECUPQ at a distribution of 50%. Furthermore, a very poor version (ECUPQ 10 bits, 50%) was added as an anchor signal. Eight listeners participated in the test. The results are presented in Figure 3, where the points represent the medians of each method, across all subjects and all excerpts, and the error bars depict the 25 and 75 percent ranges of the total response distribution.

We see that the ECUSQ scheme performs slightly better than the ECUPQ scheme with an optimal bit distribution. By changing the distribution by only 10 % a considerable drop in performance occurs for the ECUPQ scheme. We conclude from this experiment that finding the optimal bit distribution in ECUPQ is crucial for obtaining acceptable perceptual performance, misadjusting the distribution by even a few percents can lead to significant perceptual performance loss. Secondly, we can see that the ECUSQ method offers almost transparent quality at 14 bits per sinusoidal component.

4. REFERENCES

[1] H. Purnhagen, "Advances in parametric audio coding," in *Proc. 1999 IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics,*, (New Paltz, New York, USA), pp. W99-1-W99-4, 1999.

[2] T.S. Verma and T.H.Y. Meng, "A 6 kbps to 85 kbps scalable audio coder," in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.*, vol. II, (Istanbul, Turkey), pp. 887-880, 2000.

[3] R. Vafin and W.B. Kleijn, "Entropy-constrained polar quantization and its application to audio coding," accepted for *IEEE Trans. Speech Audio Processing*, 2003.

[4] P.E.L. Korten, J.Jensen and R.Heusdens "High rate spherical quantization of sinusoidal parameters," in *Proc. 12th European SP. Conf.*, (Vienna , Austria), pp. 1757-1760, September 2004.

[5] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens. "A new psycho-acoustical masking model for audio coding applications," in *Proc. ICASSP 2002*, (Orlando, Florida, USA), pp 1805-1808, May 2002.

[6] R. M. Gray and D. L. Neuhoff, Quantization. *IEEE Trans. Information Theory*, 44(6): 2325–2383, October 1998.

[7] S. P. Lloyd, Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28:129–137, 1982.

[8] R.P. Westerlaken, "High-resolution quantisation of sinusoidal parameters using a perceptual distortion measure," M.S. thesis, Delft University of Technology, Delft, The Netherlands, June 2004, Technical Report TR-200405.

[9] R. Heusdens, R. Vafin and W.B Kleijn, "Sinusoidal modelling using psychoacoustic-adaptive matching pursuits", in *IEEE Signal Processing Letters*, 9(8):262-265, August 2002.

[10] P. Prandoni, M. Goodwin, and M. Vetterli, "Optimal Time-Segmentation for Signal Modeling and Compression", in *Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc.*, (Munich, Germany), pp. 2029-2032, 1997.