

PARAMETRIC AUDIO CODING BASED ON ADAPTIVE SIGNAL MODELS

P. Vera-Candeas¹, N. Ruiz-Reyes¹, M. Rosa-Zurera², P.J. Reche-Lopez¹, J.C. Cuevas-Martinez¹

¹Electronics Department, Polytechnical School, University of Jaén
C/ Alfonso X el Sabio, 28, 23700 - Linares, Jaén, Spain

phone: +34 953 026554, fax: +34 953 026508, email: {pvera, nicolas, pjreche, jccuevas}@ujaen.es

²Signal Theory and Communications Department, Polytechnical School, University of Alcalá
Ctra. Madrid-Barcelona Km. 33,600, 28871 - Alcalá de Henares, Madrid, Spain
phone: +34 91 8856690, fax: +34 91 8856699, email: manuel.rosa@uah.es

ABSTRACT

This paper deals with the application of adaptive signal models for parametric audio coding. The matching pursuit algorithm is used for extracting sinusoidal components and transients in audio signals. The resulting residue is perceptually modelled as a noise like signal. First of all, an accurate detection of transients in audio signals is required. When a transient is detected, psychoacoustic-adapted matching pursuits are accomplished using a wavelet-based dictionary followed of an harmonic one. Otherwise, matching pursuit is applied only to the harmonic dictionary. This multi-part model (Sines + Transients + Noise) is successfully applied for audio coding purposes, assuring high perceptual audio quality at low bit rates (lower than 24 kbps for most of the CD-quality one channel audio signals considered for testing).

1. INTRODUCTION

Parametric coding of audio signals has become a popular tool for representing audio signals at very low bit rates [1, 2, 3]. A wide range of audio signals intuitively fit into the three-part model of Sines, Transients and Noise. Transients describe drum hits and the stacks of many instruments, sines describe signal components that have a distinct pitch, and noise often describes the rest of the signal that is neither sinusoidal nor transient. This model consists of three parts that work together and complement each other to form a complete and robust signal model, which makes possible a highly optimized audio compression scheme.

However, because each part of the model assumes an underlying structure to the signal, modelling errors are inevitable (i.e. if a sinusoidal model represents a signal onset, the attack becomes smeared in time, resulting in a pre-echo). To alleviate model mismatch problems, the three part of the model work together and operate in series. First, transients are modelled and removed, leaving a residual signal. Then, sinusoids are modelled and removed, leaving a noise-like signal for the noise model. As such, each model captures signal components that are coherent to its underlying assumptions.

The classical sinusoidal or harmonic model has been applied with success for the purpose of coding speech signals [4]. This model comprises an analysis-synthesis framework that represents a signal as the sum of a set of sinusoids (partials) with time-varying frequencies, phases, and amplitudes. A large number of methods have been proposed for estimating the parameters of the sinusoidal model. Estimation of parameters is typically accomplished by peak picking the

Short-Time Fourier Transform (STFT) [4]. Usually, analysis by synthesis is used in order to verify the detection of every spectral peak.

On the other hand, transients extraction is useful for those parts of audio signals with sharp attacks, because sinusoidal and noise models cannot represent them efficiently. In [3, 5, 6, 7] different approaches for transient modelling are presented.

The three-part signal model is completed with a noise model for noise-like signals. Noise modelling has seen attention in the literature. LPC based schemes are the subject of much research. Another promising noise model has perceptual roots in that it uses energy on an Equivalent Rectangular Bandwidth (ERB) scale [8]. In this paper the three-part signal model is completed with a wavelet-based noise model.

This paper proposes an efficient, accurate and flexible multi-part model for wide-band speech and audio coding. The matching pursuit algorithm is used in order to iteratively select the functions that best match the current audio frame for representing transients and sinusoids. Sinusoids are modelled using sets of complex exponential functions, while transients are modelled using sets of wavelet functions. The matching pursuit algorithm operates with both sinusoids and wavelet functions.

2. MATCHING PURSUIT

The matching pursuit algorithm was introduced by Mallat and Zhang in [9]. So as to explain the basic ideas concerning this algorithm, let's suppose a linear expansion approximating the analyzed signal $x[n]$ in terms of functions $g_i[n]$ chosen from a over-complete dictionary $D = \{g_i ; i = 0, 1, \dots, L\}$. The L elements of the dictionary span C^L and are restricted to have unit norm.

The problem of choosing the functions $g_i[n] \in D$ that best approximate the analyzed signal $x[n]$ is computationally very complex. The matching pursuit is a greedy iterative algorithm that offers a sub-optimal solution, where the two-norm is used as the approximation metric because its mathematical convenience. When a well-designed overcomplete dictionary is used in matching pursuit, the nonlinear nature of the algorithm leads to compact signal-adaptive parametric models [9].

At the first iteration of matching pursuit, the atom $g_i[n]$ which gives the largest inner product with the analyzed signal $x[n]$ is chosen. The contribution of this vector is then subtracted from the signal and the process is repeated on the residue. At the m -th iteration, the residue is:

$$r^m[n] = \begin{cases} x[n] & m=0 \\ r^{m+1}[n] + \alpha_{i(m)} \cdot g_{i(m)}[n] & m \neq 0 \end{cases} \quad (1)$$

where $\alpha_{i(m)}$ is the weight associated to the optimum atom $g_{i(m)}[n]$ at the m -th iteration, and $i(m)$ the dictionary index of the optimum atom chosen at the m -th iteration.

By computing the orthogonal projections of the residue $r^m[n]$ on elements $g_i[n] \in D$, the weight associated to each dictionary element at the m -th iteration is achieved:

$$\alpha_i^m = \frac{\langle r^m[n], g_i[n] \rangle}{\langle g_i[n], g_i[n] \rangle} = \frac{\langle r^m[n], g_i[n] \rangle}{\|g_i[n]\|^2} = \langle r^m[n], g_i[n] \rangle \quad (2)$$

The l^2 norm of $r^{m+1}[n]$ can be expressed as:

$$\|r^{m+1}\|^2 = \|r^m\|^2 - |\langle r^m, g_i \rangle|^2 = \|r^m\|^2 - |\alpha_i^m|^2 \quad (3)$$

which is minimized by maximizing $|\alpha_i^m|^2 = |\langle r^m, g_i \rangle|^2$.

Therefore, the optimum atom $g_{i(m)}$ at the m -th iteration is obtained as follows:

$$g_{i(m)} = \arg \min_{g_i \in D} \|r^{m+1}\|^2 = \arg \max_{g_i \in D} |\alpha_i^m|^2 \quad (4)$$

It is simply equivalent to choosing the atom whose inner product with the signal has the highest value.

The computation of correlations $\langle r^m[n], g_i[n] \rangle$ for all $g_i[n] \in D$ at each iteration is highly computational consuming. As derived in [9], this computation effort can be substantially reduced using an updating formula based on equation (1). The correlations at the m -th iteration are given by:

$$\langle r^{m+1}[n], g_i[n] \rangle = \langle r^m[n], g_i[n] \rangle - \alpha_{i(m)} \cdot \langle g_{i(m)}[n], g_i[n] \rangle \quad (5)$$

where the only new computation required for the correlation updating procedure refers to the cross-correlation term $\langle g_{i(m)}[n], g_i[n] \rangle$, which can be pre-calculated and stored, once overcomplete set D has been determined.

3. THE PROPOSED PARAMETRIC AUDIO CODER

The proposed parametric audio coder is defined with three meaningful components:

- Transient modelling using energy-adaptive matching pursuit with a dictionary of wavelet functions.
- Sinusoidal modelling using psychoacoustic-adaptive matching pursuit with a dictionary of complex exponentials.
- Residue modelling as a noise like signal.

Figure 1 shows the encoder stage of the proposed parametric audio coder.

The proposed audio coder extracts from the input audio signal a set of different parameters to be sent to the decoder. These parameters represent the information provided by the three-part model (Sines + Transient + Noise). They are quantified using psycho-acoustical information to ensure that decoded signals are perceptually identical to the original ones.

Before transient modelling, transient detection is required. Our transient detector is based on sudden energy change detection. Besides, an adaptive tiling of the time axis is required to achieve a right performance of the proposed audio coder. We have used the algorithm proposed in [10].

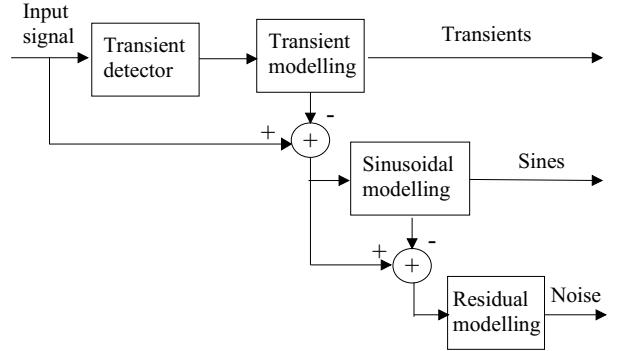


Figure 1: Block diagram of the encoder stage

3.1 Transient modelling

We propose using matching pursuits with a dictionary of orthogonal wavelet functions for transient modelling. The overcomplete dictionary D is made up with those functions which give rise to the J -depth full Wavelet-Packet (WP) decomposition, being $M_{WP} = J \cdot N$ the WP dictionary size, and N the frame length. The inner products of the signal with the wavelet-based atoms in set D lead to all the wavelet coefficients that can be considered in the J -depth full WP tree. These coefficients can be identified using three indexes, $\{i, j, k\}$, which indicate the sub-band at a given decomposition depth, the decomposition depth and the delay, respectively. The wavelet coefficients at the m -th iteration of matching pursuit and the wavelet-based atoms can be expressed as follows:

$$\alpha_{\{i,j,k\}}^m = \langle r^m[n], g_{\{i,j,k\}}[n] \rangle \quad (6)$$

$$g_{\{i,j,k\}}[n] = g_{\{i,j\}}[n - 2^j k] \quad (7)$$

According to (5), the only necessary correlations to implement the matching pursuit are $\langle x[n], g_{\{i,j,k\}}[n] \rangle$ and $\langle g_{\{i_1,j_1,k_1\}}[n], g_{\{i_2,j_2,k_2\}}[n] \rangle$. The first ones are obtained from the WP transform of $x[n]$, while correlations between atoms are pre-calculated and memory stored. These cross-correlations are formulated in [7] when wavelet-based dictionaries built from orthonormal wavelets are used, which results in:

$$\langle g_{\{i_1,j_1,k_1\}}[n], g_{\{i_2,j_2,k_2\}}[n] \rangle = \begin{cases} \delta[k_2 - k_1] & i_1 = i_2, j_1 = j_2 \\ 0 & i_2 \neq \lfloor \frac{i_1}{2^{j_1-j_2}} \rfloor \\ g_{\{i,j,k_1\}}[k_2] & i_2 = \lfloor \frac{i_1}{2^{j_1-j_2}} \rfloor \end{cases} \quad (8)$$

where $j = j_1 - j_2$ and $i = ((i_1))_{2^j}$. Therefore, according to (8), the iterative procedure to update correlations requires impulsive responses of the synthesis WP tree branches to be stored [7].

3.2 Sinusoidal modelling

For sinusoidal modelling, we propose using matching pursuits with a dictionary of windowed complex exponential functions, instead of a set of windowed sinusoidal functions,

in order to reduce the computational complexity. Using windowed complex exponential sets, only the frequency of every exponential function must be determined, which involves a significant reduction of the dictionary size [11]. The functions that belong to the considered set can be expressed as follows:

$$g_i[n] = S \cdot w[n] \cdot e^{j \frac{2\pi i}{L} n}, \quad i = 0, \dots, L \quad (9)$$

The constant S is selected in order to obtain unit-norm functions, $w[n]$ is the N -length analysis window, and $L+1$ the number of frequencies within the dictionary. Amplitude, frequency and phase are the three parameters that define each extracted tone by the sinusoidal model.

The implemented matching pursuit algorithm for sinusoidal modelling is psychoacoustic-adaptive as in [12]. According to this approach, the extracted tone at each iteration is the perceptually most important one. Psychoacoustic-adaptive matching pursuits [12] define a perceptual distortion measure as

$$\|PD_i\|^2 = \int_0^1 \hat{a}(f) |(\widehat{w[n]}(\alpha_i^m g_i[n]))(f)|^2 df \quad (10)$$

where $\hat{\cdot}$ indicates the Fourier transform operation, $w[n]$ is a window defining the signal segment, and \hat{a} is the inverse of the masking threshold, which is computed on the basis of the reconstructed signal that changes with each iteration.

In our implementation, the perceptual distortion measure in equation (10) is slightly modified by integrating directly along the bark band scale, which results in a complexity reduction.

Another improvement to be remarked is the contribution of a perceptual stopping criteria. In [12], the masking threshold is initialized to the threshold in quiet at the first iteration, being added the mask due to each extracted tone at the remaining iterations. However, the noise-to-tone masking is not considered at all, which makes unfeasible a perceptual stopping procedure. Instead, we initialize the masking threshold to the threshold in quiet plus the mask owing to the noise (as computed in MPEG-2 layer III), which assures that only perceptually meaningful tones are extracted, if the algorithm is halted when the extracted tone at a given iteration is below the masking threshold.

3.3 Residual modelling

After sinusoidal and transient components are extracted from the original signal, the residue is considered to be a noise like signal. For audio applications, psychoacoustic phenomena have to be incorporated in the parametric noise model. For noise perception, the exact shape of the magnitude spectrum is not as crucial as the energy at each critical band. According to this principle, the ERB noise modelling is proposed in [8]. In our approach, the ERB model is approximated by the Discrete Wavelet Transform (DWT). This noise model adheres to the general idea of filtering the output of a white noise source to approximate the shape of the residue in the frequency domain. In this case, DWT dictates the form of the filter bank, performing a dyadic partition in frequency, which plays a central role in many aspects of perception.

The proposed noise model is composed of two stages: analysis and synthesis. The DWT-based analysis stage di-

vides each frame into $J+1$ wavelet bands (being J the decomposition depth), and estimates their energy. For the l -th frame, the energy of the r -th wavelet band is found as:

$$E_r^l = \sum_{m \in \beta_r} |X^l(m)|^2 \quad (11)$$

where β_r contains the indexes of the r -th wavelet band, and $X^l(m)$, $m \in \beta_r$, represents the wavelet coefficients of the r -th wavelet band for the l -th frame.

The energy parameters approximates a power spectrum with piecewise constant energy according to the DWT filter bank. These parameters are used for the DWT-based synthesis stage. In the synthesis stage the wavelet coefficients are initialized to white noise using each band energy to control its respective gain, which results in the synthesized noise.

Subjective listening tests pointed the necessity of improving the time characteristics of the synthesized noise in order to avoid spreading effects. LPC filtering has been included in the proposed noise model to achieve a time shaping of the synthesized noise. We have applied an Auto-Regressive all poles model with 4 poles as maximum. The number of poles in the model is given by the prediction gain. A lattice structure is adopted to achieve an efficient quantization of the AR model information included in our noise modelling approach.

4. RESULTS AND DISCUSSION

To assess the performance of the proposed audio coder, we have obtained some subjective and objective results. The configuration parameters are: 32-coefficient Daubechies filters and 4-level full WP decompositions ($J = 4$) for transient modelling, 4096 frequencies ($L = 4096$) within the dictionary for sinusoidal modelling, and 32-coefficient Daubechies filters and 9-level depth for DWT in noise modelling. Six music samples considered hard to encode have been used. They are 15 seconds-length CD-quality one channel audio signals. Special attention has been paid to signals with impulsive energy bursts, which are extremely susceptible to the presence of 'pre-echoes', and we have made sure that the chosen set of audio source material covers a wide variety of signals. It is listed in table 1.

Table 1: SOURCE MATERIAL USED IN THE TEST

TEST ITEM	DESCRIPTION
si01	Harpsichord
si02	Castanets
sm01	Bagpipes
sm02	Glockenspiel
sc01	Trumpet solo and orchestra
sc02	Orchestra piece

4.1 Objective results

The resulting binary rates obtained with the proposed parametric audio coder are presented in table 2. It contains the partial bit rates resulting for the synthetic signals obtained from sinusoidal, transient and residual modelling and the final bit rates resulting for the decoded audio signals (in kbit/s).

In order to illustrate the performance of the proposed audio coder, let's consider an audio frame with an impulsive

Table 2: BIT RATES

Item	Tones	Transients	Residue	Decoded signals
si01	15.7	3.43	4.49	23.62
si02	8.78	7.33	6.62	22.73
sm01	16.17	2.44	4.71	23.32
sm02	5.34	2.98	3.15	11.47
sc01	20.57	0.62	2.59	23.78
sc02	19.72	0.41	5.17	25.3

energy burst. Figure 2(a) represents the original audio signal, while figures 2(b) and 2(c) represent the synthesized transient and sinusoidal components, respectively, when they are modelled using the above described approaches. Finally, figure 2(d) shows the noise-like residual signal. It can be observed that the synthetic signal in figure 2(b) properly represents the sharp attack in the original one.

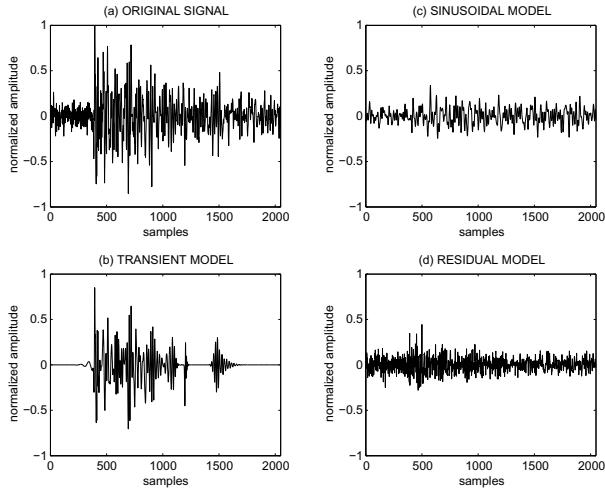


Figure 2: Synthetic signals obtained from transient, sinusoidal and residual modelling.

4.2 Subjective results

The subjective tests have been performed on headphones under the A-B-C rule using the sequences shown in table 1. The A-B-C methodology, known as a triple-stimulus double blind test with hidden reference, is recommended by ITU-R in the BS. 1116-1 recommendation. Tests have been carried out with twenty trained listeners, and the results are shown in table 3.

Table 3: SUBJECTIVE RESULTS UNDER THE ITU-R BS.1116-1 RECOMMENDATION.

Test Items	Orig. MOS	Decoded MOS	Δ MOS
si01	4.98	4.45	0.53
si02	5	4.07	0.93
sm01	4.99	4.51	0.48
sm02	5	4.13	0.87
sc01	5	4.40	0.6
sc02	5	4.28	0.72

5. CONCLUSIONS

This paper deals with parametric audio representation for audio coding. The used model considers the audio signal composed of three kinds of components: sinusoidal, transients and noise like components. For estimating the parameters of the sinusoidal and transient models, matching pursuit with dictionaries of complex exponentials and wavelet functions, respectively, is used. A novel wavelet-based noise modelling is applied for residue modelling, which is completed with LPC filtering to achieve Time Noise Shaping (TNS). The proposed audio coder achieves nearly transparent coding at very low bit rates (lower than 24 kbit/seg). Hence, our coder is a good proposal for audio coding applications at very low bit rates, as Internet streaming.

REFERENCES

- [1] S. Levine and J. Smith, "A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications", *105th AES Convention*, preprint 4781, September 1998.
- [2] T.S. Verma, "A perceptually based audio signal model with application to scalable audio compression", *PhD thesis*, Standford University, 1999.
- [3] A.C. Den Brinker, E.G.P. Schuijers and A.W.J. Oomen, "Parametric coding for high quality audio", *112th AES Convention*, Preprint 5554, May 2002.
- [4] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. 34, no. 4, 1986, pp. 744-754.
- [5] J. Nieuwenhuijse, R. Heusdens, and E.F. Deprettere, "Robust exponential modeling of audio signals", *Proc. ICASSP-98*, vol. 6, May 1998, pp. 3581-3584.
- [6] T. Verma and T. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio", *Proc. ICASSP-98*, May 1998.
- [7] P. Vera, N. Ruiz, M. Rosa, D. Martinez and F. Lopez, "Transient Modeling by Matching Pursuits with a Wavelet Dictionary for Parametric Audio Coding", *IEEE Signal Processing Letters*, To appear.
- [8] M. Goodwin, "Residual modelling in music analysis-synthesis", *Proc. ICASSP-96*, vol. 2, May 1996, pp. 1005-1008.
- [9] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Trans. on Signal Processing*, vol. 41, no. 12, December 1993, pp. 3397-3415.
- [10] N. Ruiz, M. Rosa, F. López and P. Vera, "New algorithm for achieving an adaptive tiling of the time axis for audio coding purposes", *Electronic Letters*, vol. 80, April 2002, pp. 434-435.
- [11] M. M. Goodwin, *Adaptive Signal Models. Theory, Algorithms and Audio Applications*, 1998, Kluwer Academic Publishers.
- [12] R. Heusdens, R. Vafin and W. B. Kleijn, "Sinusoidal Modelling using Psychoacoustic-Adaptive Matching Pursuits", *IEEE Signal Processing Letters*, vol. 9, no. 8, August 2002.