Adaptive Signal Models for Wide-Band Speech and Audio Compression

Pedro Vera-Candeas[1], Nicolás Ruiz-Reyes[1], Manuel Rosa-Zurera[2], Juan C. Cuevas-Martinez[1], and Francisco López-Ferreras[2]

[1] Electronics and Telecommunication Engineering Department, University of Jaén Polytechnic School, C/ Alfonso X el Sabio 28, 23700 Linares, Jaén, Spain
{pvera,nicolas,jccuevas}@ujaen.es
[2] Signal Theory and Communications Department, University of Alcalá Polytechnic School, 28871 Alcalá de Henares, Madrid, Spain
{manuel.rosa,francisco.lopez}@uah.es

Abstract. This paper deals with the application of adaptive signal models for parametric speech and audio compression. 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. 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 speech and audio coding purposes, assuring high perceptual quality at low bit rates (close to 16 kbps for most of the signals considered for testing).

1 Introduction

Parametric coding of audio signals has become a popular tool for representing these signals at very low bit rates [1–3]. 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. To alleviate model mismatch problems, the three part of the model 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

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3523, pp. 571–578, 2005.

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] 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 [7]. 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 [8]. 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 $D = \{g_i \ ; \ i = 0, 1, \ldots, L\}$. The L elements of the dictionary span C^L and are restricted to have unit norm.

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:

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

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 that atom.

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

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

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

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

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:

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

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 [8], 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:

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

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 Wide-Band Speech and Audio Coder

The proposed parametric wide-band speech and 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 a complex exponentials.
- Residue modelling as a noise like signal.

Figure 1 shows the encoder stage of the proposed parametric wide-band speech and audio coder.

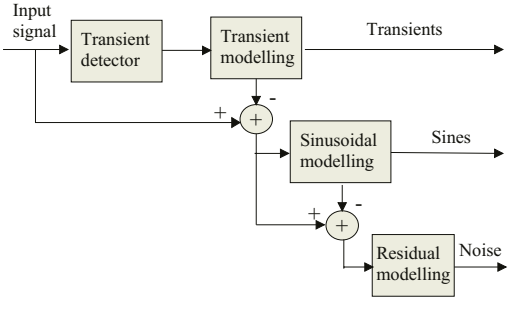

Fig. 1. Block diagram of the encoder stage.

The proposed wide-band speech and 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 [9].

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:

$$\alpha^m_{\{i,j,k\}} = \langle r^m[n], g_{\{i,j,k\}}[n] \rangle \tag{6}$$

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

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 [6] when wavelet-based dictionaries built from orthonormal wavelets are used, which results in:

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

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 [6].

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 [10]. The functions that belong to the considered set can be expressed as follows:

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

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 [11]. According to this approach, the extracted tone at each iteration is the perceptually most important one. Psychoacoustic-adaptive matching pursuits [11] define a perceptual distortion measure as

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

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

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

3.3 Residual Modelling

After sinusoidal and transient modeling, the residue is considered to be a noise like signal. For audio applications, psychoacoustic phenomena have to be incorporated into the 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 [7]. In our approach, the ERB model is approximated by the Discrete Wavelet Transform (DWT). 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 divides 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:

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

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 wide-band speech and 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. Twelve music samples considered hard to encode have been used. They are 15 seconds-length CD-quality one channel speech and 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 source material covers a wide variety of signals.

4.1 Objective Results

The resulting binary rates obtained with the proposed wide-band speech and audio coder are presented in table 1. 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 signals (in kbits/s).

In order to illustrate the performance of the proposed wide-band speech and audio coder, let's consider an audio frame with an impulsive 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.

4.2 Subjective Results

The subjective tests have been performed on headphones under the A-B-C rule using the twelve 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 2.

Table 1. Bit rates.

Item	Description	Tones	Transients	Residue	Decoded signals
es01	Suzanne Vega	12.14	0.98	3.34	16.52
es02	German male speech	12.48	0.78	3.37	16.69
es03	English female speech	13.94	0.97	3.00	17.98
si01	Harpsichord	11.73	0.25	2.54	14.60
si02	Castanets	11.84	4.30	2.38	18.61
si03	Pitch pipe	8.21	0.15	3.50	11.90
sm01	Bagpipes	9.22	0.17	3.75	13.20
sm02	Glockenspiel	3.76	0.67	2.36	6.85
sm03	Plucked strings	13.94	0.14	2.80	16.93
sc01	Trumpet solo and orchestra	13.00	0.45	2.87	16.38
sc02	Orchestra piece	12.76	0.20	2.25	15.26
sc03	Contemporary pop	15.60	0.21	2.85	18.73

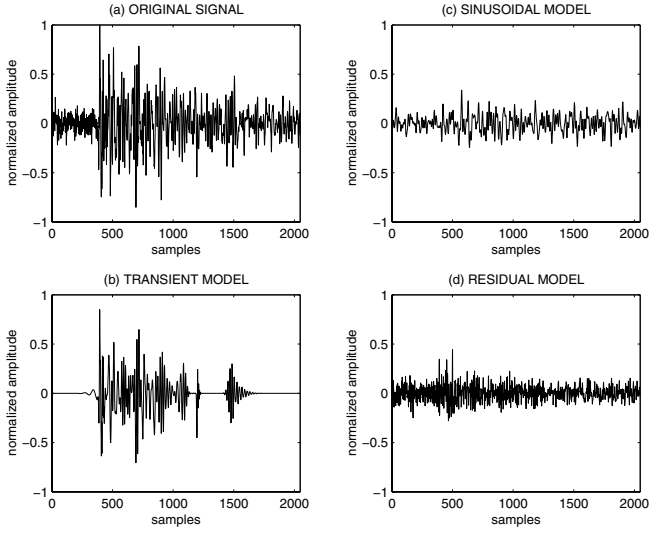

Fig. 2. Synthetic signals obtained from transient, sinusoidal and residual modelling.

5 Conclusions

This paper deals with parametric representation for wide-band speech and audio coding. The used model considers the speech and audio signals 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 wide-band speech and audio coder achieves nearly transparent coding

Table 2. Subjective results under the ITU-R BS.1116-1 recommendation.

Test Items	Orig. MOS	Decoded MOS	ΔMOS
es01	5.00	4.19	0.81
es02	5.00	4.02	0.98
es03	5.00	4.12	0.88
si01	4.97	4.63	0.34
si02	5.00	4.55	0.45
si03	5.00	4.33	0.67
sm01	4.99	4.51	0.48
sm02	4.98	4.68	0.30
sm03	4.97	4.75	0.22
sc01	5.00	4.40	0.60
sc02	5.00	4.28	0.72
sc03	5.00	4.33	0.67

at very low bit rates (close to 16 kbit/seg). Hence, our coder is a good proposal for audio coding applications at very low bit rates, as Internet streaming.

References

1. Levine, S., Smith, J.: A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications, *105th AES Convention*, preprint 4781 (1998).
2. Verma, T.S.: A perceptually based audio signal model with application to scalable audio compression, *PhD Thesis*, Standford University (1999).
3. Den Brinker, A.C., Schuijers, A.G.P., Oomen, A.W.J.: Parametric coding for high quality audio, *112th AES Convention*, Preprint 5554 (2002).
4. McAulay, R., Quatieri, T.: Speech Analysis/Synthesis Based on a Sinusoidal Representation, *IEEE Trans. Acoustic, Speech and Signal Processing* **34** 4 (1986) 744-754.
5. Nieuwenhuijse, J., Heusdens, R., Deprettere, E.F.: Robust exponential modeling of audio signals, *Proc. ICASSP-98* **6** (1998) 3581-3584.
6. Vera-Candeas, P., Ruiz-Reyes, N., Rosa-Zurera, M., Martinez-Muñoz, D., Lopez-Ferreras, F.: Transient Modeling by Matching Pursuits with a Wavelet Dictionary for Parametric Audio Coding, *IEEE Signal Processing Letters* **11** 3 (2004) 349-352.
7. Goodwin, M.: Residual modelling in music analysis-synthesis, *Proc. ICASSP-96* **2** (1996) 1005-1008.
8. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries, *IEEE Trans. on Signal Processing* **41** (1993) 3397-3415.
9. Ruiz, N., Rosa, M., López, F., Vera, P.: New algorithm for achieving an adaptive tiling of the time axis for audio coding purposes, *Electronic Letters* **80** (2002) 434-435.
10. Goodwin, M.M.: Adaptive Signal Models. Theory, Algorithms and Audio Applications, *Kluwer Academic Publishers* (1998).
11. Heusdens, R., Vafin, R., Kleijn, W.B.: Sinusoidal Modelling using Psychoacoustic-Adaptive Matching Pursuits, *IEEE Signal Processing Letters* **9** 8 (2002).