ATS: A System for Sound Analysis Transformation and Synthesis Based on a Sinusoidal plus Critical-Band Noise Model and Psychoacoustics

Juan Pampin
Center for Digital Arts and Experimental Media (DXARTS), University of Washington
pampin@u.washington.edu

Abstract

ATS is a spectral modeling system based on a sinusoidal plus critical-band noise decomposition. Psychoacoustic processing informs the system's sinusoidal tracking and noise modeling algorithms. Perceptual Audio Coding (PAC) techniques such as Signal-to-Mask Ratio (SMR) evaluation are used to achieve perceptually accurate sinusoidal tracking. SMR values are also used as a psychoacoustic metric to determine the perceptual relevance of partials during analysis data post-processing. The system's noise component is modeled using Bark-scale frequency warping and sub-band noise energy evaluation. Noise energy at the sub-bands is then distributed on a frame-by-frame basis among the partials resulting in a compact hybrid representation based on noise modulated sinusoidal trajectories. This paper presents the most relevant aspects of the ATS system.

1 Introduction

ATS belongs to a family of spectral modeling systems that could be traced back to the seminal work on speech analysis/synthesis by R. J. McAulay and T. F.Quatieri [1], the high-resolution spectrum analysis research done by J. O. Smith in the mid-1980s [2] with its later extension into PARSHL [3], and SMS, the ground breaking work by X. Serra based on deterministic plus stochastic decomposition of sound [4].

The main contribution of ATS to the Spectral Modeling field is the introduction of psychoacoustic and Perceptual Audio Coding (PAC) techniques both in the sinusoidal and noise analysis phases. Based on psychoacoustic information, the two parts of the model (sinusoids and noise) are blended by means of critical-band noise energy distribution among the partials. In ATS, a partial is no longer a sinusoidal component but a hybrid component that contains a time-varying noise element modulated by a sinusoidal trajectory.

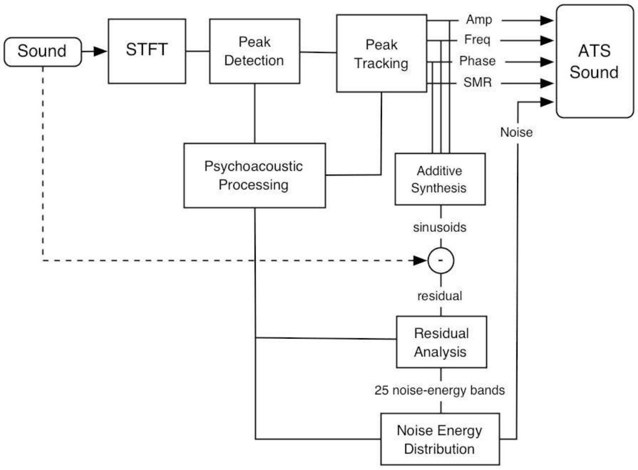

Figure 1: ATS system block diagram

2 The ATS System

As discussed in the introduction, ATS is based on a sinusoidal plus critical-band noise model, this means that the system assumes that the sound to be analyzed can be represented accurately with those components, i.e. time-varying sinusoids modulated by critical-band noise. The following sub-sections describe each part of the system as shown on Fig. 1.

2.1 STFT Analysis

The ATS system uses the Short Time Fourier Transform (STFT) as its main analysis tool. The sound is analyzed using overlapping time windows and taking the STFT on each window.

2.2 Peak Detection

After converting the STFT data into polar form, a peak detection algorithm is performed on the dB magnitude spectrum, which issues a set of relevant spectral peaks for the corresponding analysis frame.[1]

2.3 Psychoacoustic Processing

At this point, the system performs psychoacoustic processing on the analysis frame. This involves masking curve evaluation and computation of the Signal-to-Mask Ratio (SMR) for each spectral band. SMR information (in dB SPL scale) is stored together with the corrected frequency, magnitude and phase values for each peak issued by the peak detection algorithm. PAC techniques used in this process are discussed in Section 3.

2.4 Peak Tracking

The next step involves frame-to-frame tracking of peaks. The goal here is to connect peaks that "belong" to the same spectral trajectory. In other words, we want to extract sinusoidal partials out of the time-changing spectral peak texture. Both frequency and SMR information are used in this process. This algorithm will be described in Section 4.

2.5 Residual

Once traced, sinusoidal trajectories can be extracted from the analyzed sound to obtain a time-domain residual. This residual should consist of a noise-like signal representing what couldn't be modeled with sinusoids. This process is achieved by performing additive synthesis of the sinusoidal trajectories (using phase information) and subtracting the result from the original sound in the time domain [2].

2.6 Residual Analysis

After isolating the two modeling components (sinusoids and residual), noise analysis is performed on the residual. The result of this step (which is described in detail in Section 5) is a set of 25 time-varying noise energy bands.

2.7 Noise Energy Distribution

The main goal of the ATS system is creating a parametric representation of a spectral model. In other terms, a representation capable of an infinite number of transformations by manipulation of a perceptually meaningful set of parameters. For this reason a compact model is desired, where both the sinusoidal and noise components are blended into a single representation. This is performed by distributing the noise energy in each critical band to the partials on a frame-by-frame basis. This process of "hybridizing" partials is discussed in Section 6.

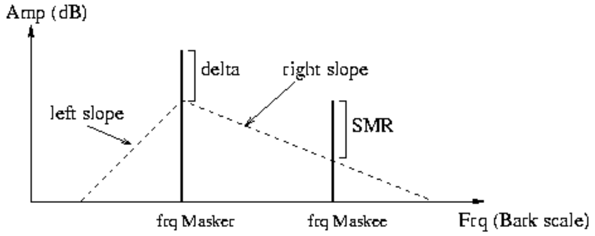

Figure 2: Signal-to-Mask Ratio (SMR) evaluation using a simple masking model.

3 Psychoacoustics

ATS uses psychoacoustic processing for both its sinusoidal tracking and noise modeling algorithms. During this processing phase, PAC methods are used to gather psychoacoustic information from the STFT analysis data. The use of PAC methods doesn't have the goal of data compression but that of the computation of a psychoacoustic metric. This metric consists mainly on the evaluation of masking curves using a simple masking model and the computation of the SMR of each spectral component.

3.1 Masking Model

A simple masking model is used to evaluate the SMR of each spectral component. This model requires frequency to be warped into Bark scale. The main reason for this is that masking profiles can be modeled with linear equations in the Bark domain, also the 24 steps of the Bark scale correspond to the critical bands of hearing which are used for the noise analysis phase.

The masking model consists of three components: 1) the difference between the level of the masker and the masked threshold (called delta, typically -10 dB SPL); 2) the masking profile slope towards lower frequencies (typically around 27 dB/Bark); 3) the masking profile slope towards high frequencies (typically around 15 dB/Bark) [5].

[1] ATS' peak detection algorithm is, in general terms, the same as the one presented in [3] and [4].

[2] This procedure is in general terms the same as the one presented in [1] and [4].

3.2 SMR Evaluation

Using a recursive algorithm, masking curves are evaluated considering each peak as a masker for the rest of the peaks (maskees) present in a given frame. By the end of this process, each peak will be assessed a SMR value in dB SPL. Fig. 2 illustrates the masking evaluation process.

4 Peak Tracking

Peaks issued by the peak detection algorithm need to be connected and translated into spectral trajectories. This process involves the evaluation of the possible candidates to continue trajectories on a frame-by-frame basis. This is done using tracks that keep information of average values for each of the trajectory parameters. The length of the tracks is adjustable and has to be tuned depending on the characteristics of the analyzed sound. The best candidate peak for a particular spectral track would be the one that meets the following criteria:

$$Min\left\{\Delta = \frac{|P_{freq}-T_{freq}|+\alpha|P_{SMR}-T_{SMR}|}{1+\alpha}\right\}$$

where P_{freq} is the candidate peak frequency, and P_{SMR} its SMR, T_{freq} is the track frequency, and T_{SMR} its SMR both averaged over the track length (typically 3 frames). α is a coefficient controlling how much the SMR deviation affects the computation of Δ. The algorithm recursively finds the best candidate (i.e. the one with the minimum Δ) for each track, continuing the spectral trajectories. New tracks get created from orphan peaks (the ones that were not incorporated to any existing tracks), and tracks which couldn't find continuing peaks are set to sleep[3].

4.1 SMR Continuation

The use of the SMR continuation as a parameter for the peak tracking process is based upon psychoacoustic temporal masking phenomena [5]. Conceptually, we assume that masking profiles of stable sinusoidal trajectories can only evolve at slow rate (no sudden changes). This is true for analysis performed with hop sizes between 10 and 50 milliseconds, which is comparable to the average duration of pre- and post-making effects [5].

5 Post-processing

Once the sinusoidal trajectories have been traced post processing is performed on the data to fix trajectory defects that might occur during the tracking. Such defects mainly consist of time gaps in otherwise coherent long trajectories or short trajectory segments that are not perceptually relevant.

5.1 Fixing Time Gaps

Trajectory gaps shorter than a certain number of frames (a parameter of the post-processing algorithm, typically 3 frames) are fixed by interpolation. The partial data at the edges of the gap feeds appropriate interpolation functions for each of the parameters, which get filled frame by frame.

5.2 Short Segment Pruning

Short trajectory segments may appear in the analysis, especially in transient regions. Short segments with average SMR below a certain threshold can be discarded without perceptual impact. This threshold is computed as a linear function of the length of the segment. The shorter the segment the larger should be its average SMR for the segment to be kept.[6]

6 Residual Analysis

As discussed in Section 2.5, a residual signal is computed taking the time-domain difference between the original sound and the sinusoidal synthesis of the spectral trajectories. This noise signal is then analyzed using the STFT to obtain the energy at each of the 25 critical bands [5]. Residual analysis is performed at the same frame rate of the sinusoidal analysis.

6.1 Frequency Warping

The first step consists of converting the frequency components of the STFT analysis into Bark scale. Once this is performed, each step in the the scale corresponds to a critical band of hearing and magnitudes in each band can be used to evaluate its noise energy.

6.2 Noise Energy Evaluation

Noise energy in each critical band is evaluated in the following way:

$$E[i] = \frac{1}{K}\sum_{k=0}^{K-1}|X(k)|^2$$

where i is the band number (0 to 24), K is the number of bins of the STFT Xpresent in the band. The algorithm evaluates the noise energy at each step of the Bark scale.

[3] This part of the algorithm is very similar to the one presented in [4].

7 Noise Energy Distribution

The residual analysis algorithm issues 25 values of noise energy for each analysis frame. This time-varying noise energy bands perceptually represent the evolution of the residual in the frequency domain. The following step consists on distributing the energy in each critical band to the partials, creating hybrid sinusoidal plus noise trajectories.

7.1 Partial Noise Energy Computation

For each analysis frame, the algorithm recursively detects partials in each critical band and distributes the band's noise energy among them. The noise energy for each partial is computed as:

$$P_E[j] = \frac{P_{SMR}[j]*E[i]}{S[i]}$$

where j is the partial number, P_E its noise energy, P_{SMR} its SMR, and i the band number to which the partial belongs. The denominator S is computed as:

$$S[i] = \sum_{j=0}^{J-1} P_{SMR}[j]$$

where J is the total number of partials in band i.

8 Synthesis

Hybrid partials are synthesized using an extended additive synthesis technique. This technique consists of two parts: the generation of a normalized sinusoidal component (without amplitude scaling) and the generation of a noise component. For a particular partial P, the two parts (called P_{sin} and P_{noi} respectively) are calculated as:

$$P_{sin} = \cos{(2\pi P_{freq})},\ P_{noi} = \text{randi}(P_E, P_{bw})$$

where P_{freq} is the partial's time-varying frequency, *randi* is a linear-interpolation random number generator producing new random values at a P_{bw} rate which directly depends on P_{freq}. Random values generated by *randi* are scaled using the time-varying energy of the partial, P_E. Finally, the partial is synthesized by the addition of the amplitude scaled sinusoidal component plus the modulated noise component:

$$P_{synth} = P_{amp}P_{sin} + P_{noi}P_{sin}$$

where P_{synth} is the synthesized partial (for all frames), P_{amp} the partial's time-varying amplitude, P_{noi} its noise component, and P_{sin} its sinusoidal component. This procedure is repeated for all partials and results summed to generate the synthesized sound Z:

$$Z = \sum_{j=0}^{J-1} P_{synth}[j]$$

where J is the total number of partials in the analysis.

9 Conclusion

ATS introduces psychoacoustic and Perceptual Audio Coding (PAC) techniques to the field of spectral modeling. Providing a perceptual metric (SMR), psychoacoustic processing informs both the sinusoidal and noise analysis parts of the system. The two components of the model (sinusoids and noise) are blended by means of critical-band noise energy analysis and distribution, creating hybrid components that contain a time-varying noise element modulated by a sinusoidal trajectory. This compact parametric model is capable of an infinite number of transformations by the manipulation of a small set of perceptually meaningful parameters. The ATS system has been evaluated and tested with a wide variety of sounds, generating results perceptually equivalent to the analyzed sources. ATS has been used as a composition tool by many composers of diverse esthetic orientations proving a wide application scope and flexibility [7].

References

[1] McAulay, R. J., and T. F. Quatieri. 1986. "Speech Analysis/Synthesis based on a Sinusoidal Representation." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34(4): 744-754.

[2] Smith, J. O., and B. Friedlander. 1984. "High Resolution Spectrum Analysis Programs." Memo 5466-05. Systems Control Technology, Palo Alto, California.

[3] Smith, J. O., and X. Serra. 1987. "PARSHL: An Analysis/Synthesis program for Nonharmonic Sounds based on a Sinusoidal Representation." *Proceedings of the 1987 International Computer Music Conference.* San Francisco: Computer Music Association.

[4] Serra, X. 1989. "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition." PhD diss., Stanford University.

[5] Zwiker, E. and H. Fastl. 1990. *Psychoacoustics Facts and Models.* Springer, Berlin, Heidelberg.

[6] García, G., and J. Pampin. 1999. "Data Compression of Sinusoidal Modeling Parameters Based on Psychoacoustic Masking" *Proceedings of the 1999 International Computer Music Conference.* Beijing: Computer Music Association.

[7] Pampin, J. 1999. "ATS: a Lisp Environment for Spectral Modeling" *Proceedings of the 1999 International Computer Music Conference.* Beijing: Computer Music Association.