PERCEPTION-BASED RESIDUAL ANALYSIS-SYNTHESIS SYSTEM

Yin H Lam and Robert W Stewart

Signal Processing Division, Dept. of Electrical & Electronic Engineering
University of Strathclyde, Glasgow, G1 1XW, UK

ABSTRACT

The paper describes a residual analysis-synthesis system which exploits the human perception mechanism on temporal varying signals using Zwicker's three dimensional excitation-critical-band-rate time pattern as the framework. Temporal information is retrieved using a linear predictive analysis on the discrete cosine transformed signal and critical band intensity information is obtained by using non-uniform filter banks. The system is characterized by high frequency resolution and good time resolution. Novel phase prediction and phase correction techniques are employed to eliminate any boundary discontinuities between two time frames. Experimental results illustrate that high quality residual signals can be reproduced using a few parameters regardless of the temporal characteristics of the signal.

1. INTRODUCTION

Despite the success of contemporary sinusoidal models such as the sinusoidal transformation system [1] and the spectral modeling synthesis [2] in reproducing high quality quasi-periodic signals, the reproduction of stochastic components or the so called residual signals are not of good quality and also invariably require a large number of parameters. Noise driven source filter models in conjunction with phase randomization have been proposed to synthesize the stochastic components by convolving white noise with time-varying frequency-shaping filters represented by a short time spectral envelope [2] or the equivalent rectangular bands energy of the auditory system [3]. Boundary discontinuities can then be removed using window overlap-add techniques [2] [3]. In spite of their capability in modeling stochastic signals using a few parameters, they are unable to reproduce highly time localized events. Waveform matching algorithms [4] [5] have been proposed recently to handle the transients but at the cost of using more parameters. Strategies inspired from solutions to the pre-echo problem in the audio coding literature such as automatic time segmentation scheme [6] reduce the amount of pre-echo but at the cost of coding efficiency and poor frequency resolution. Also the degradation due to the window overlap-add analysis-synthesis process is still not adequately addressed [6].

In this paper, a Perception-Based Residual Analysis-Synthesis (PBRAS) system is proposed to synthesize the stochastic components regardless of the temporal characteristics by exploiting the limits of human hearing. The three dimensional excitation-critical-band-rate time pattern (ECBRTP) in Zwicker's Temporal Excitation/Loudness Pattern model [7] forms the basic framework of our system. ECBRTP is realized using non-uniform filter banks to extract the critical band intensity. Temporal variation in each critical band is retrieved by performing a linear predictive (LP) analysis on the discrete cosine transform (DCT) of the critical band signal. High frequency resolution is retained as no switching between a long and short frame size is required. A coherent loss compensation scheme is introduced to replace the undesirable conventional window overlap analysis process and signals are reconstructed using the ECBRTP information described by the LP filter coefficients and the critical band intensity. The number of synthesis parameters are minimized from a perceptual point of view as the LP filter orders can be adjusted to adapt to the varying temporal characteristics in each critical band. Instead of using phase randomization, randomly spaced frequency components assignment in the frequency domain is employed to produce the “noise sensation”. Boundary artifacts are virtually completely removed using novel phase prediction and phase correction techniques and exploiting the forward masking characteristics of human hearing. This therefore eliminates the additional window overlap-add processes often required for conventional residual signal synthesis system.

2. TEMPORAL EXCITATION/LOUDNESS PATTERN MODEL

Our proposed PBRAS system is based on Zwicker's temporal excitation/loudness pattern model [7] which aims to quantify the subjective loudness sensation and estimate a hearing equivalent psychoacoustical value perceived by the human auditory system. The model is based on the assumption that the auditory system behaves as if it contains a bank of bandpass filters commonly called critical bands with continuous overlapping center frequencies and that the frequency selectivity can be approximated by subdividing the intensity of sound into each filter band. Apart from spectral distribution, the model further points out that temporal information is essential to the human hearing perception. The time dependent excitation value in each critical band or the so called excitation-critical-band-rate time pattern is the most essential intermediate value that best illustrates the transformation of physically defined audio stimuli into the loudness sensation. Such three dimensional patterns are responsible for the different hearing sensations owing to different audio stimuli and hence serve as the basic framework of our perception-based residual analysis-synthesis system on noise modeling.

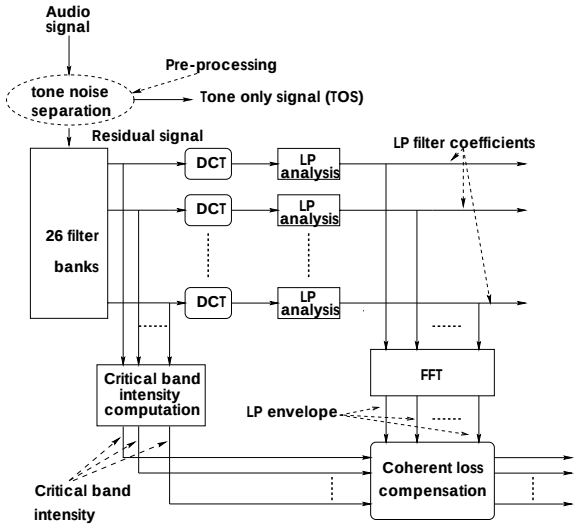

Figure 1: The block diagram of the perception-based residual analysis system.

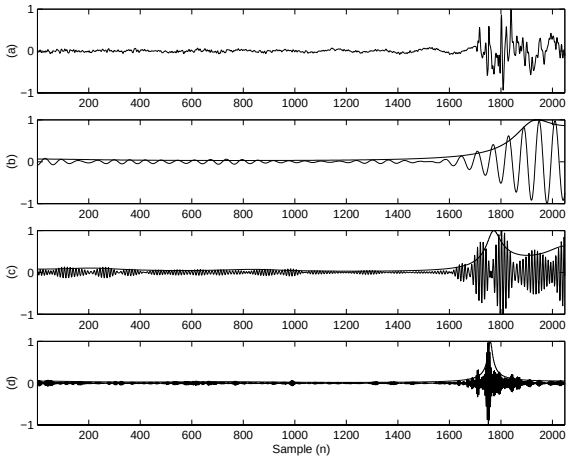

Figure 2: (a) The original full band residual waveform of a drum beat. (b), (c) and (d) show the signal components in critical band 7, 18, and 23 and their LP temporal envelope using LP filter order, 3, 7, 5 respectively.

3. PERCEPTION BASED RESIDUAL ANALYSIS-SYNTHESIS SYSTEM

3.1. Analysis

The schematic block diagram of the residual analysis system is shown in Figure 1. Peak picking algorithms and frequency domain subtraction [1] [2] are employed in the pre-processing block to separate the tonal and noise components from each other. The filter banks, the temporal information extraction block and the coherent energy loss compensation block form the three main components of the residual analysis system.

3.1.1. Filter banks

Non-uniform filter banks are used to decompose the residual signal into 26 critical band components to match the frequency selectivity of human hearing. As the filter banks are only for analysis purposes, properties such as sharp transition bands and high stop-band attenuation rather than perfect reconstruction are preferred. An FIR filter using a Hamming window design method was employed in our system. As well as converting the signal into 26 time-frequency components, the filter output can also provide a direct measure of the critical band intensity in each critical band.

3.2. Discrete Cosine Transform and Linear Predictive Analysis

The underlying concept behind our temporal information retrieval approach comes from the time-frequency duality principle: an LP analysis in the time domain can produce spectral envelope information and hence by simple duality a frequency domain analysis can provide temporal envelope information. Time-frequency duality has been recently exploited in Temporal Noise Shaping (TNS) in MPEG-2 Advanced Audio Coding [8] and Transient Modeling Synthesis (TMS) [5]. Our system differs from their schemes in both the key aim and the algorithmic approaches. TNS was introduced in MPEG-2 to adapt the quantization error to the temporal shape of the input signal by quantizing the linear prediction residual in the frequency domain rather than the spectral components. The TMS approach, however, attempts to synthesize the transients in the time domain using sinusoidal modeling in the frequency domain. Our aim is to retrieve the psychoacoustic three dimensional excitation-critical-band-rate time pattern to exploit the perceptual limits of the human auditory system. The temporal information extraction block consists of 26 Discrete Cosine Transforms (DCT) which map the filter bank output signals to real value samples in the frequency transform domain and followed by the linear predictive (LP) analysis [9]. The output parameters are 26 sets of LP filter coefficients with the individual filter order adapted to the temporal variation within each critical band. Using an appropriate filter order selection criteria, signals can be described by a minimal set of parameters from the perceptual point of view. Figure 2 gives an example of adaptive order allocation on a drum attack. The original full-band normalized residual is shown in Figure 2(a). Figure 2(b), (c), (d) show the the normalized signal components from critical band 7, 18 and 23 and their LP temporal envelope with LP filter order 3, 7 and 5 respectively. In order to achieve the best compromise between the number of LP parameters and the fit of the LP envelope to the temporal variation in each critical band, the prediction error, a by-product of the Levinson-Durbin [9] algorithm is used to determine the optimum orders and the LP filter coefficients.

3.2.1. Coherent loss compensation

A coherent loss compensation scheme is used in conjunction with the randomly spaced frequency components noise modeling in the synthesis side of PBRAS to compensate the reduction in coherent

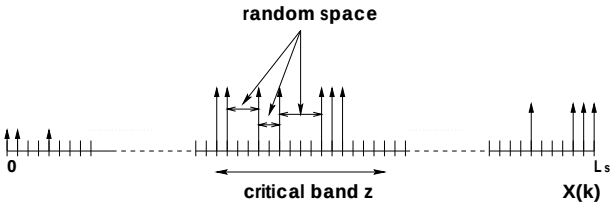

Figure 3: Randomly spaced frequency components assignment using FFT bins.

gain in analyzing non-stationary signals using stationary signal assumption [10]. The unavailability of a temporal window in the analysis process is the main cause for the coherent loss when analyzing a non-stationary signal. In PBRAS system, the temporal envelope is readily obtained by calculating the Fourier transform of the impulse response of the LP filter.

3.3. Synthesis

3.3.1. Residual synthesis using random frequency components

The ear loses its frequency pattern recognition power and produces noise sensation when too many randomly spaced frequency components fall on the basilar membrane at the same time [11]. Therefore in our system the signals are reconstructed by modeling the noise in each time segment using large number of sinusoidal components randomly spaced in the frequency domain. The number of frequency components can be decided beforehand. As shown in Figure 3, the implementation is equivalent to the assignment of amplitude and phase information to the FFT (Fast Fourier Transform) bin. Phases are randomly assigned at the beginning of the first time frame. The same amplitude are assigned to frequency components within same critical band and they can be directly derived from the critical band intensity and the number of components within the critical band.

3.3.2. Frequency domain additive synthesis and convolution

An IFFT procedure [12] is used to synthesize the time domain signal $x(n), 1 \leq n \leq L_s$, where n is the time domain index and L_s is the synthesis segment length. Temporal information which is represented by the LP temporal envelope can be incorporated by multiplying the individual LP envelope with the time domain signal corresponding to each critical band and followed by a summation in the time domain. As time domain multiplication can be implemented using frequency domain convolution, the final FFT spectrum $X(k)$ corresponding to the time domain operation is equivalent to

$$X(k) = \sum_{z=1}^{26} a^z * S^z, \qquad 0 \leq k \leq 2 \times L_s - 1 \tag{1}$$

$$S^z = \sum_{k=0}^{2 \times L_s - 1} A_k \exp j\theta_k \tag{2}$$

where a^z is the set of LP filter coefficients in the critical band z and $A_k = 0, k < k_{zl}, k > k_{zu}$ or $k_{zl} \leq k \leq k_{zu}$ when bin k is not assigned. k_{zl} is the first FFT bin and k_{zu} the last FFT bin corresponding to critical band z. The time domain signal $x(n)$ is obtained from the inverse FFT of $X(k)$. The incorporation of the LP temporal envelope together with the coherent loss compensation scheme prevents the undesirable energy redistribution owing to the conventional window-overlap add processes [10].

3.3.3. Boundary phase prediction and correction

Boundary artifacts will occur if the phases are randomized from frame to frame and window overlap-add synthesis will be required to reduce the undesirable edge effect. However, as the random frequency assignment produces adequate noise sensation without phase randomization, we can predict the phase of each frequency component f_k at the beginning of the next frame $i+1$ to reduce the boundary artifacts.

$$\theta_k^{i+1}(0) = 2\pi f_k L_s + \theta_k^i(0) \tag{3}$$

Any slight discontinuity due to abrupt signal changes can be further removed by phase correction technique. A phase correction procedure is introduced by shifting the waveform until the signal components at the boundary are continuous. Phase shifting is equivalent to adding $\delta\theta_k^{i+1}$ to $\theta_k^{i+1}(0)$ where $\delta\theta_k^{i+1}$

$$\delta\theta_k^{i+1} = \arccos\left(\frac{A_k^i}{A_k^{i+1}\cos\theta_k^{i+1}(0)}\right) - \theta_k^{i+1}(0) \tag{4}$$

Hence, phase prediction and phase correction remove most boundary artifacts without using overlap-add synthesis except when a very weak signal follows a strong signal. However, such artifacts are usually inaudible as they are often masked by the forward temporal masking [7].

4. RESULTS

Six different audio sources were used in the subjective listening tests. They were taken directly from compact disc recordings with sampling rate $44100kHz$, 16 bits resolution and mono. The ITU-R five-point impairment scale [13] was used to assess the subjective performance of PBRAS using fixed frame size of 2048 and a stochastic synthesis scheme (SS) with 512 frame size using window overlap-add noise-driven source filter model and steady state excitation pattern when compare to the tone-only signals (TOS) which resulted from the removal of the stochastic components of the original signals in the tone-noise separation block. 512 frame size is used for SS as this give approximately the same number parameters used in PBRAS. The number of parameters required for PBRAS, the Perceptual Distance (PD) and the Signal-to-Noise Ratio are shown in Table 1. Perceptual Distance is defined as the absolute difference between ITU-R impairment scores of the original and the synthetic signals. They are shown here instead of the ITU-R impairment score because the original were hidden in the tests and some subjects failed to identify the original music. From the results we note that PBRAS has better PD performance than both SS and TOS. It is noted the PD performance of TOS is the worst of all, although the objective performance SNR is the best. In fact,

		Perceptual Distance			SNR (dB)		
Source	No. Parameters	PBRAS	SS	TOS	PBRAS	SS	TOS
saxophone + drum	110	1.00	1.33	1.84	15.51	15.38	18.12
castanet + string instruments	127	0.82	1.25	2.68	17.20	16.05	17.92
classical guitar	101	0.56	0.74	0.99	18.37	18.60	20.59
drum + base guitar + jazz	127	1.18	1.80	3.33	11.14	10.90	13.95
piano + bell	99	0.23	0.55	0.58	25.18	24.81	27.05
piano + rain + thunderstorm	103	0.45	1.00	1.58	20.97	20.84	24.20

Table 1: The performance of PBRAS, SS and TOS using different audio sources.

PBRAS also excelled SS and TOS among subjects who correctly identified the original music segments in terms of the ITU-R impairment scores. However, the current PBRAS synthetic signals still have some distance from transparency for polyphonic signal. Detailed investigation revealed that the degradation is mainly due to the presence of undesirable residual tonal components which the pick-peaking algorithms failed to separate.

Two conclusion can be reached from the results. First, the PBRAS synthetic signal resembles the original music more closely. Second, residual modeling is useful in narrowing the subjective difference due to the removal of stochastic components.

5. CONCLUSIONS AND FUTURE RESEARCH EFFORT

Our perception-based residual analysis-synthesis system (PBRAS) has demonstrated that the excitation-critical-band-rate time pattern is effective in describing the subjective hearing sensation of different stochastic audio stimuli with varying temporal characteristics. LP analysis in the DCT domain and the use of non-uniform filter banks are shown to be capable of realizing the underlying psychoacoustic model with a few parameters. The introduction of coherent loss compensation scheme, phase prediction and correction strategies and the random spaced frequency components assignment ensure high quality signal reconstruction. However, experimental observations reveal the robustness of the tone-noise separation process has ultimate influence on the final subjective performance. The degraded performance of peak picking algorithms for polyphonic audio signal implies the development of more robust tone-noise separation algorithm. Audio demonstration files of PBRAS are available at
$http://www.spd.eee.strath.ac.uk/users/vicky/audiofile.html$

6. REFERENCES

[1] R. J. McAulay, and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744-754, 1986.

[2] X. Serra, and J. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, 1990, vol. 14, no. 4, pp, 12-24, 1990.

[3] M. Goodwin, "Residual modeling in music analysis-synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2, pp. 1005-1008, 1996.

[4] Y Ding, and X. Qian, "Processing of musical tones using a combined quadrature polynomial-phase sinusoidal and residual (QUASAR) signal model," *Journal of Audio Engineering Society*, vol. 45, no. 7/8, pp. 571-584, 1997.

[5] T. S. Verma, S. N. Levine and T. H. Y. Meng, "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," *International Conference of Computer Music, Greece*, 1997.

[6] P. Prandoni, M. Goodwin, and M. Vetterli, "Optimal time segmentation for signal modeling and compression," *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 3, pp. 2029-2032, 1997.

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

[8] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchsi, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," *Journal of Audio Engineering Society*, vol. 45, no. 10, pp. 789-814, 1997.

[9] T. Parsons, *Voice and Speech Processing*, New York: McGraw-Hill, 1987.

[10] Y. H. Lam, and R. W. Stewart, "Overlap-add free musical noise analysis-synthesis", *IEEE Workshop on Signal Processing Systems (SiPS'98)*, Cambridge, MA, October, 1998.

[11] J. G. Roederer, *Introduction to the Physics and Psychophysics of Music*, Springer-Verlag, New York, Heidelberg. Berlin, 1975.

[12] E. B. George, and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model", *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 389-406, 1997.

[13] ITU-R BS. 1116, *Methods for the subjective assessment of small impairment in audio systems including multichannel sound systems*, Geneva, Switzerland, 1994.