SOUND TEXTURE MODELLING WITH LINEAR PREDICTION IN BOTH TIME AND FREQUENCY DOMAINS

Marios Athineos and Daniel P.W. Ellis

Dept. of Electrical Engineering, Columbia University, New York 10027
{marios,dpwe}@ee.columbia.edu

ABSTRACT

Sound textures—for instance, a crackling fire, running water, or applause—constitute a large and largely neglected class of audio signals. Whereas tonal sounds have been effectively and flexibly modelled with sinusoids, aperiodic energy is usually modelled as white noise filtered to match the approximate spectrum of the original over 10-30 ms windows, which fails to provide a perceptually satisfying reproduction of many real-world noisy sound textures. We attribute this failure to the loss of short-term temporal structure, and we introduce a second modelling stage in which the time envelope of the residual from conventional linear predictive modelling is itself modelled with linear prediction in the *spectral* domain. This cascade time- and frequency-domain linear prediction (CTFLP) leads to noise-excited resyntheses that have high perceptual fidelity. We perform a novel quantitative error analysis by measuring the proportional error within time-frequency cells across a range of timescales.

1. INTRODUCTION

Sound modelling is concerned with capturing a signal's important information in a simplified parameter space. Models for speech are frequently motivated by the goal of minimizing data size for ease of coding and transmission. In computer music, signal models can provide parameters that permit interesting and coherent modifications such as pitch and timing variations.

The classic source-filter model for voice takes a periodic or noisy excitation (for vowels and fricatives respectively) and passes it through a time-varying filter simulating the vocal tract. Linear predictive modelling [1], in which the filter is an autoregressive, all-pole model, has been particularly successful because of its low complexity analysis and synthesis. In musical signals, sinusoidal models of individual Fourier components have been very successful. Realism is improved by adding a random-noise background, filtered to match the residual left after sinusoid modelling [2, 3]. Quality can be improved still further by separate detection and modelling of brief energy bursts known as transients [4, 5].

In this paper, we look at a third class of sounds we call *sound textures* that are distinct from speech and music, and which call for their own specially-designed models. Although a rigorous definition is elusive, examples of textural sounds include applause, running water, rainfall, fire, babble, and machinery. Like their visual namesakes, textures should have an indeterminate extent (duration) with consistent properties (at some level), and be readily identifiable from a small sample. Texture analysis and synthesis is an interesting challenge, and has potential applications in general sound recognition, virtual reality synthesis, and abstract-level coding. Prior work on sound textures includes [6, 7].

Many of the sounds we have collected as textures are *noisy* (i.e. without strong, stable periodic components) and *rough* (i.e. amplitude modulated in the 20-200 Hz range [8]). Under conventional speech or music models, a sound like applause would most likely be represented as a sequence of spectral estimates for frames of 10-30ms duration. However, a resynthesis consisting of white noise excitation filtered to match these estimates loses much of the timbral texture of the original, indicating that the noise is a poor substitute for the ideal analysis residual. An obvious difference lies in the temporal distribution of energy within the frame: Our textural sounds are often composed of many individual brief events—rain splashes, hand claps, or fire crackles—which persist as concentrated *microtransient* noise bursts in the residual.

In the next section, we present an extended model that captures extra structure from the residual, to permit noise-excited resyntheses with excellent perceptual fidelity. The performance of this model is evaluated in section 3 through a novel error criterion that measures local signal similarity over a wide range of time scales. In section 4, we discuss issues raised by this model as well as some potential applications. Section 5 contains our conclusions.

2. CASCADE TIME-FREQUENCY LINEAR PREDICTION (CTFLP) MODEL

Our goal is a parametric modelling representation that can preserve the character of textures, supporting high quality noise-excited resynthesis as well as transformations such as time scale modification (TSM). We model textures as rapidly-modulated noise by using two linear predictors in cascade. The first, operating in the time domain, captures the spectral envelope, whereas the second, operating in the frequency domain, captures the temporal envelope of the texture. The synthesis step recreates the original texture by feeding random noise through the estimated filters in series. There are several advantages of this approach over a more conventional deterministic plus stochastic model: It is not oriented around sinusoids which are usually weak or absent from textures; Transients or microtransients are not treated any differently from the rest of the signal, reducing complexity and avoiding artifacts from transient detection and separation; and the flexible representation of the temporal envelope is particularly effective in applications such as time-scale modification.

2.1. Frequency Domain Linear Prediction

The part of the model that is responsible for the accurate representation of the temporal structure of rough textures is the frequency domain linear prediction (FDLP), a concept first introduced by Herre and Johnston [9]. Dubbed temporal noise shaping (TNS),
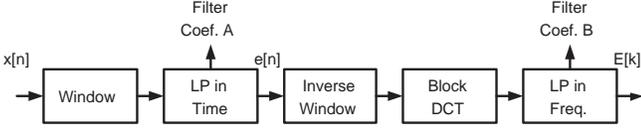
Fig. 1. CTFLP Analysis block diagram.
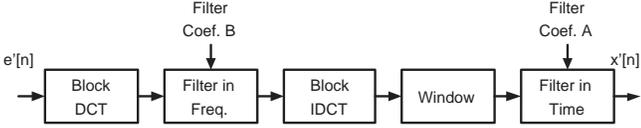
Fig. 2. CTFLP Synthesis block diagram.

its principal application was in the elimination of pre-echo artifacts associated with transients in perceptual audio coders. Using D*PCM coding of the frequency domain coefficients, the authors showed that coding noise could be shaped to lie under the temporal envelope of the transient.

FDLP is the frequency domain dual of the well-known time domain linear prediction (TDLP). In the same way that TDLP estimates the power spectrum, FDLP estimates the temporal envelope of the signal, specifically the square of its Hilbert envelope,

$$e(t) = \mathcal{F}^{-1} \left\{ \int \tilde{X}(\zeta) \cdot \tilde{X}(\zeta - f) d\zeta \right\} \qquad (1)$$

i.e. the inverse Fourier transform of the autocorrelation of the of the single sided (positive frequency) spectrum $\tilde{X}(f)$. We use the autocorrelation of the spectral coefficients to predict the temporal envelope of the signal.

2.2. CTFLP Analysis and Synthesis

The sequence of operations in our cascade time-frequency linear predictive (CTFLP) analysis is shown in figure 1. First, a frame of the original signal is multiplied by a time window. This is a necessary preprocessing step of the autocorrelation method for TDLP [10]. After TDLP spectral estimation, the whitened residual is obtained by passing the signal through an inverse filter. Before proceeding to the FDLP, the original temporal envelope is restored by undoing the initial window, to avoid wasting the modelling power of the FDLP on the window envelope. We use the discrete cosine transform (DCT) to obtain the single sided spectrum, and estimate the temporal envelope using FDLP. The final residual is flat both in its spectral and temporal envelopes.

CTFLP synthesis is illustrated in figure 2. The analysis residual is not particularly close to Gaussian white noise, as measured by the skewness and kurtosis of its cumulative distribution function. Perceptually, however, it is practically indistinguishable from a simple noise sequence. To permit coherent excitation in our 50% frame overlaps, we use a time-domain random sequence as the starting point for our resynthesis. The remainder of the resynthesis simply reverses the analysis procedure, first by filtering the DCT spectrum of the noise using the coefficients extracted in the FDLP, then using the time-domain filter from the analysis TDLP to reim-

pose the broad spectral structure of the original frame. Frames are overlapped to recreate a continuous signal.

3. EVALUATION

We used the CTFLP procedure described above to produce noise-excited resyntheses of our small collection of sound textures. Informal listening tests show cascade modelling to be extremely successful at preserving the character of sounds, capturing both the spectral and temporal characteristics of rough noisy textures such as the fizz of pouring soda out of a bottle. (Sound examples are available on our website[1].) The technique has greatest success with sounds that include both broadband noise and densely-packed microtransients. Such sounds are very difficult to represent by methods that detect and separate transients from the rest of the residual.

We performed a quantitative analysis of the difference between noise-excited resyntheses using the new CTFLP approach and a conventional TDLP scheme. To be fair, both algorithms used the same windows and LP methods, and we equalized the number of parameters in each system. For our main results, we used 40 time-domain and 10 frequency-domain poles in the CTFLP case, and 50 time-domain poles in the simple TDLP scenario.

To reveal the variation of modelling error with temporal scale, we devised an error metric based on the short time Fourier transform (STFT) magnitude. For a given temporal window length (simultaneously defining the spectral resolution), the STFT magnitude in every time-frequency cell is calculated for both original and resynthesized (TDLP or CTFLP) signals with 50% window overlap. This is repeated over a range of error analysis window lengths between 1 ms and 1 second.

We define the per-cell mean proportional magnitude (MPM) error as:

$$E_{MPM} = \frac{1}{NM} \sum_{n=0}^{N-1} \sum_{k=0}^{M-1} \left| \frac{|X(n,k)| - |X_{LP}(n,k)|}{|X(n,k)| + \varepsilon} \right| \qquad (2)$$

where $X(n,k)$ is the STFT cell at time step n and frequency bin k. Since the noise based resynthesis is a random process, we calculate the error by averaging over around 50 resynthesis realizations. The ε in the denominator reduces the dominance of cells in which original signal is almost zero; it is set to 10% of the average cell magnitude.

In figure 3 we compare the absolute error for CTFLP and TDLP of fast applause (appl) and people yelling (yell), with the default arrangement of 50 poles per frame in the TDLP, and 40+10 poles in CTFLP. Firstly, we notice that for the applause signal (round symbols), the new CTFLP (hollow symbols) achieves much lower error than conventional TDLP (filled symbols) for error analysis windows smaller than the window of 22ms used in the LP modelling (the "modelling window"). This is as expected, since the main point of CTFLP is to preserve the temporal energy distribution below the level of the frame, lost in TDLP. Note also that for error analysis windows longer than the modelling window, CTFLP exhibits slightly worse error than TDLP; at this scale, the error analysis is looking only at the spectral match within each frame, and the temporal modulation imposed by the FDLP stage will slightly distort the minimum mean-squared error fit achieved by pure TDLP.

Looking at the hollow and filled squares, we see that for a sound such as yelling (multiple pitched voices) that contains few
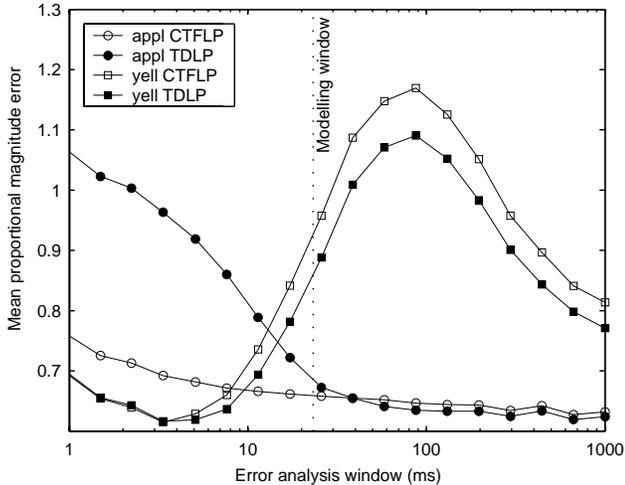
Fig. 3. Mean Proportional Magnitude Error as a function of the error analysis window length. Two sound textures are shown: applause ("appl", circles), which is noisy and rich in microtransients, and yelling ("yell", squares), which is largely tonal and has few transients. Error curves are plotted for both conventional TDLP (filled symbols) and the proposed CTFLP (hollow symbols).

Fig. 4. MPM Error ratio of CTFLP to TDLP, as a function of the error analysis window, for a range of sounds: yelling ("yell") and laughter ("laug") are more tonal and smooth, whereas soda being poured from a bottle ("bott"), fast applause ("appl") and rapid typewriter noise ("type") are more noisy and rough, and thus show the advantage of CTFLP at short timescales.

transients and mostly harmonic energy, the time-domain modelling of CTFLP confers no advantage and is consistently worse than the TDLP, which can use its entire pole budget for a slightly better spectral model. However, once the error analysis window is large enough to distinguish the separate harmonics in the original (i.e above about 10ms), neither noise-excited resynthesis can perform particularly well.

Figure 4 presents the same results, this time combining the errors under each model into a single curve per sound example showing the ratio of the MPM error of CTFLP to TDLP, and including results for several other sounds. Points above the dotted horizontal line where the error ratio is 1 achieve a smaller error under conventional TDLP. For rough textures like the typewriter (type), applause (appl) and the bottle (bott) we see that CTFLP clearly outperforms TDLP in the low error analysis window region, paying a small penalty at the larger time scales. For non-rough, tonal textures like yelling (yell) or laughter (laug) we see that the error ratio is consistently greater than 1 (CTFLP performing worse).

In figure 5, we examine the behavior of MPM error for a fixed error analysis window of 5ms, showing the results as a function of the modelling window length. Two parameter strategies are used: fixed poles per frame (fppf), in which the number of poles is kept constant for every modelling frame at 50 for TDLP and 40+10 CTFLP, and hence the total number of parameters increases as the frames become shorter. In fixed poles per time (fppt), we decrease the number of poles available for shorter frames, to keep constant the total number of poles used for the whole sound. (In all cases, CTFLP uses 80% of its poles for time-domain modelling and the remainder for FDLP). We see that the variation of error with modelling window length is much smaller for CTFLP (hollow symbols), particularly under the fppt strategy. This indicates that CTFLP is relatively immune to window choice, and has no need for complex strategies such as window switching. At very short modelling windows (below the 5ms error analysis window),
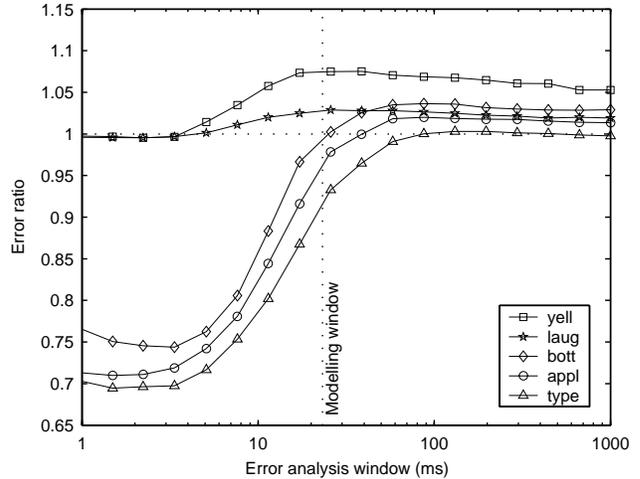
the TDLP and CTFLP approaches converge since any improved temporal resolution is hidden from this analysis. fppf achieves a lower error in this range because it has more poles available.

Figure 6 shows the effect of varying the number of frequency domain poles used in CTFLP between 1 and 20. The MPM error is expressed as a ratio to the error obtained with plain TDLP using the same number of poles. Even using just one pole in the frequency domain achieves a noticeable improvement under short error analysis windows, and as the number of FDLP poles is increased to 20 the error ratio drops consistently. At 10 poles and above, we see a spectral distortion penalty with the error ratio exceeding 1 for error analysis windows larger than the modelling window. The figure of 10 poles, used in the other results presented here, is confirmed as a good compromise between error improvement at the short time scale and minimal distortion at longer time scales.

4. DISCUSSION

The two LP models in CTFLP are essentially marginalized time and frequency envelopes used to model the full t-f distribution within each modelling frame. However, our sequential analysis is suboptimal, as shown by the worsening of error at long timescales as the number of temporal poles increases. Instead, we aim to investigate a simultaneous solution that optimizes the error for the whole analysis, which might also involve formulating FDLP in the time domain, analogously to the use of the FFT to calculate the autocorrelation at the heart of TDLP. Another aspect to the optimization is the allocation of model parameters (poles) between the time and frequency estimates; the 40/10 split used here worked well for our examples, but in general this could be varied according to the structure of the detail in individual frames. A difficulty in devising an optimal solution, however, is the poorly defined criterion of perceptual quality: No single error analysis window adequately captures the perceptually salient properties of the resynthesis, al-
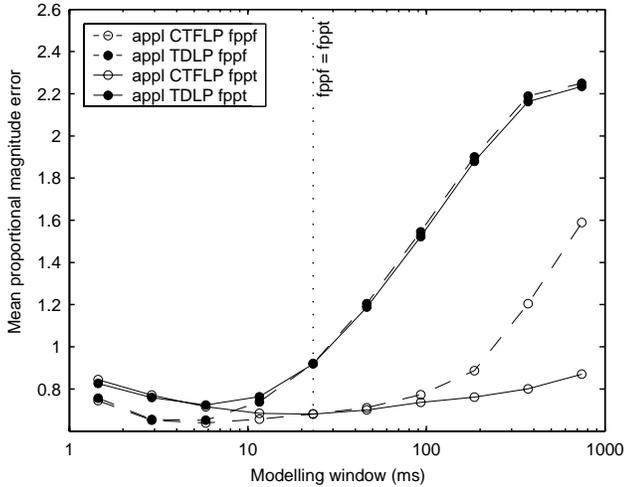
Fig. 5. MPM error for a 5ms error analysis window as a function of modelling window length, for CTFLP and TDLP, using a fixed number of poles per frame (fppf), or a fixed number of poles per unit time (fppt, where shorter frames receive fewer poles). TDLP becomes steadily worse at reproducing temporal detail as the modelling window gets longer, whereas CTFLP is better able to preserve detail at long modelling windows.

Fig. 6. Ratio of CTFLP to TDLP error as a function of the error analysis window length for a variety of FDLP pole counts. Increasing the count of FDLP poles reduces errors at time scales shorter than the modelling window, but causes significant additional error at larger time scales as it is increased to 20. Ten FDLP poles per frame emerges as a good compromise.

though a multiresolution analysis based on the properties of hearing might be better. A full-blown perceptual error metric would need to include masking effects etc. Other potential extensions of the system include a multiband variant in which the spectrum is divided prior to FDLP to estimate different temporal envelopes for different frequency bands, as mentioned in [9].

Applications of this representation include efficient encoding and synthesis of textural sounds e.g. for virtual reality systems. In such a scenario, it is desirable to employ the extensible nature of textures to generate unlimited, nonrepeating stretches of a particular texture. This requires a statistical model of synthesis parameters; we are hopeful that the CTFLP model will provide a suitable parameter space for this kind of generative sound texture model.

The current system has proved remarkably effective at time-stretching transients even up to 8 times the original length while preserving the character of the original sound and virtually eliminating time-smearing. Comparisons between the CTFLP, TDLP and phase vococder methods (available on our website) show the clear superiority of CTFLP.

5. CONCLUSIONS

Although the category of sound textures is difficult to define with precision, there are a great many everyday sounds that present difficulties to conventional modelling schemes optimized for voice and music. A relatively simple extension to model the temporal envelope of the spectral-modelling residual by using linear prediction in the spectral domain has been observed to greatly improve the perceptual fidelity of noise-excited reconstructions. An analysis constructed to measure the energy distribution error at different time scales confirms that short-term structure is preserved far better than under conventional spectral modelling, at least for sounds with dense transients and little harmonic content. This new para-

metric model holds great promise as a basis for all kinds of sound texture analysis and generation schemes.
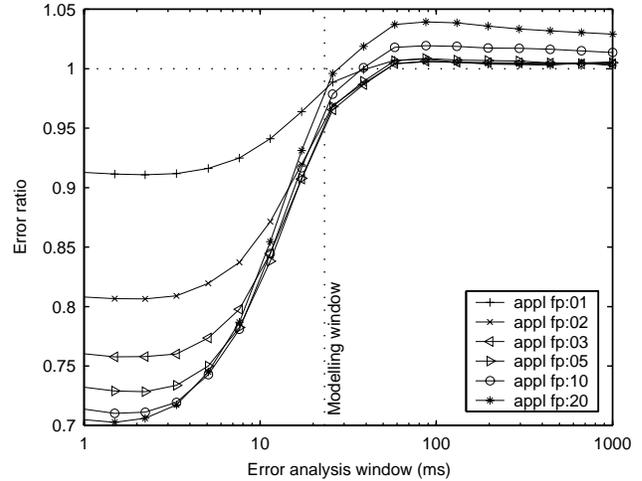
6. REFERENCES

[1] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, 1976.

[2] X. Serra, "Musical Sound Modeling with Sinusoids plus Noise," in *Musical Signal Processing*, G. De Poli et al., Ed. Swets & Zeitlinger, 1997.

[3] M. Goodwin, "Residual Modeling in Music Analysis-Synthesis," in *Proc. ICASSP*, 1996, vol. 2, pp. 1005–1008.

[4] T. Verma, S. Levine, and T. Meng, "Transient Modeling Synthesis: a flexible analysis/synthesis tool for transient signals," in *Proc. ICMC*, 1997, vol. 2, pp. 164–167.

[5] H. Thornburg and F. Gouyon, "A Flexible Analysis-Synthesis Method for Transients," in *Proc. ICMC*, 2000.

[6] N.S. Arnaud and K. Popat, "Analysis and Synthesis of Sound Textures," in *Computational Auditory Scene Analysis*, D.F. Rosenthal and H.G. Okuno, Eds., pp. 293–308. LEA, 1997.

[7] S. Dubnov et al., "Synthesizing sound textures through wavelet tree learning," *IEEE CGA*, vol. 22, no. 4, pp. 38–48, Jul/Aug 2002.

[8] E. Terhardt, "On the perception of periodic sound fluctuation (roughness)," *Acustica*, vol. 30, pp. 201–212, 1974.

[9] J. Herre and J.D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," in *Proc. 101st AES Conv.*, Nov 1996.

[10] S.M. Kay, *Modern Spectral Estimation: Theory & Application*, Prentice-Hall, 1988.