SYNTHESIS OF SOUND TEXTURES WITH TONAL COMPONENTS USING SUMMARY STATISTICS AND ALL-POLE RESIDUAL MODELING

Hyung-Suk Kim and Julius Smith

Center for Computer Research in Music and Acoustics (CCRMA)
Stanford University, USA
{hskim08|jos}@ccrma.stanford.edu

ABSTRACT

The synthesis of sound textures, such as flowing water, crackling fire, an applauding crowd, is impeded by the lack of a quantitative definition. McDermott and Simoncelli proposed a perceptual source-filter model using summary statistics to create compelling synthesis results for non-tonal sound textures. However, the proposed method does not work well with tonal components. Comparing the residuals of tonal sound textures and non-tonal sound textures, we show the importance of residual modeling. We then propose a method using auto regressive modeling to reduce the amount of data needed for resynthesis and delineate a modified method for analyzing and synthesizing both tonal and non-tonal sound textures. Through user evaluation, we find that modeling the residuals increases the realism of tonal sound textures. The results suggest that the spectral content of the residuals has an important role in sound texture synthesis, filling the gap between filtered noise and sound textures as defined by McDermott and Simoncelli. Our proposed method opens possibilities of applying sound texture analysis to musical sounds such as rapidly bowed violins.

1. INTRODUCTION

Sound *textures* are signals that have more structure than filtered noise, but, like visual textures, not all of the details are perceived by the auditory system. Saint-Arnaud [1] gives a qualitative definition of sound textures in terms of having constant long term characteristics that, unlike music or speech, do not carry a message which can be decoded. Figure 1 illustrates the relative information-bearing potential of music, speech, sound textures, and noise, showing how sound textures lie between music/speech and noise. Examples of sound textures include natural sounds such as water flowing, leaves rustling, fire crackling, or man-made sounds like the sound of people babbling, a crowd applauding or sounds of machinery such as drills. There can be textural components in musical sounds such as fast violin-bowing or guitar-string scraping.

A better understanding of sound textures can provide insights into our auditory process, and what information we extract from auditory inputs. Furthermore, such knowledge can be used to find sparse representations and applied to analysis/synthesis of environmental sounds, sound texture identification, data compression, and gap-filling.

Since Saint-Arnaud's work on sound texture, there has been a gradual increase of interest in this area and various approaches have been explored [2, 3]. One approach that has been extensively used is granular synthesis [4, 5, 6, 7, 8, 9]. In most cases, the general approach is to parse the audio during analysis, usually into sound events and background din, and then recompose the components according to a stochastic rule. The advantage of these approaches is that the original source is used for resynthesis, resulting in output quality as good as the source. This also means, however, that the method is bound by the source signal and that the methods may not be generalizable. Other approaches include applying various metrics and theories such as polyspectra [10], wavelets [11], dynamic systems [12] and scattering moments [13] to analyze sound textures.

Another approach is that of source-filter modeling. One source-filter approach is time-frequency linear predictive coding (TFLPC) also called cascade time-frequency linear prediction (CTFLP) [14, 15]. In TFLPC, time domain linear prediction, which captures the spectral content, is followed by frequency domain linear prediction, which models the temporal envelope of the residuals.

McDermott and Simoncelli [16] propose a source-filter approach using perceptual multiband decomposition, looking at the long term statistics of the multiband signal and its modulations. To evaluate the proposed model, the extracted statistics are imposed onto subband envelopes using an iterative method. The subband envelopes are multiplied with a noise signal to create the synthesized signal. An advantage of this approach is that there are no assumptions regarding the nature of the sound source, as it models how the auditory system processes the sound.

A limitation of the method proposed by McDermott and Simoncelli is that it does not work well for tonal sounds. The resynthesized results of sounds with tonal components such as wind chimes and church bells were perceived to have low realism.

Liao et al.[17] applied McDermott and Simoncelli's approach directly to a short-time Fourier transform (STFT), where marginals and subband correlations are extracted from the STFT of source signal, then iteratively imposed onto a new STFT for resynthesis.

Although there are a set of sounds that are generally accepted as sound textures, such as water flowing, fire crackling, and babble noise, there is not yet a generally quantitative definition for sound textures. Moreover, how sound texture is defined or rather defin-

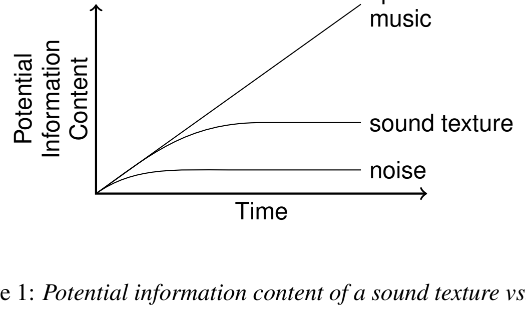


Figure 1: *Potential information content of a sound texture vs. time (from Saint-Arnaud[4])*

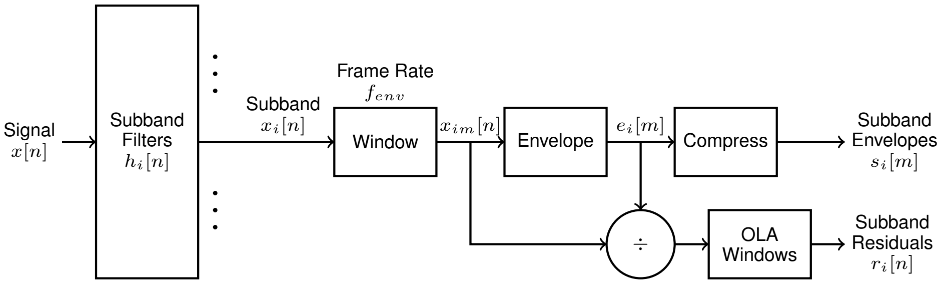


Figure 2: *Sound texture decomposition. The schematic illustrates the sound texture decomposition process for a single subband.*

ing the scope of sound textures, i.e., specifying which sounds are included in a given class of sound textures, in turn affects the approach to analyzing and synthesizing sound textures in that class.

In this paper, we limit our definition of sound textures to what can be synthesized using structured noise, that is, sounds that can be reproduced by shaping noise in a structured way. Despite the limiting definition, this approach can cover a broad range of sounds, as demonstrated by the aforementioned source-filter models. With this definition, sounds with pitch inflections, such as the sound of a baby crying or cars accelerating, will likely not fall into this category and thus will not be considered. Such a definition works well in conjunction with sines+noise synthesis [18] in which sinusoidal modeling handles any tonal components while texture classification, analysis, and synthesis can be applied to the residual signal after the tonal components are removed.

In the following sections, we examine sound textures with tonal components, compare it to non-tonal sound textures, and apply the insights gained from the comparisons to the developement of an analysis/synthesis model that includes tonal components.

2. SOUND TEXTURE DECOMPOSITION

We begin by formulating a method to decompose a sound texture into its subband sideband modulations, which we will call envelopes, and its residuals.[1] The decomposition process is illustrated in Figure 2.

The source sound texture $x[n]$ is first separated into subbands $x_i[n]$ with a subband filter bank equally spaced on an equivalent rectangular bandwidth (ERB) scale $h_i[n]$ [19]. We choose $h_i[n]$ such that its Fourier transform $H_i[k]$ satisfies,

$$\sum_i \left|H_i[k]\right|^2 = \mathbf{1}. \tag{1}$$

Thus, $\{h_i\}$ forms an FIR power-complementary filter bank [20]. The filter bank $h_i[n]$ is applied to the signal for both the analysis and synthesis steps. The analysis step gives

$$x_i[n] = h_i[n] * x[n]. \tag{2}$$

For subband $x_i[n]$, we first apply an analysis window $w[n]$ with 50% overlap on the signal at frame rate f_{env}. The length of the window is $L = 2R = 2/f_{env}$. We choose $w[n]$ to have constant overlap-add (OLA), i.e.,

$$\sum_m w^2[n + mR] = \mathbf{1}. \tag{3}$$

The window $w[n]$, like $h_i[n]$, is applied to the signal for the analysis and synthesis steps. We define the m-th windowed segment of $x_i[n]$ as

$$x_{im}[n] = w[n]x_i[n - mR]. \tag{4}$$

For each subband segment $x_{im}[n]$, we derive the uncompressed envelope of the segment $e_i[m]$ by taking the power within the windowed segment and normalizing it by the squared sum of the window $w[n]$,

$$e_i[m] = \left\{ \frac{\sum_{n=0}^{L-1}(x_{im}[n])^2}{\sum_{n=0}^{L-1}(w[n])^2} \right\}^{\frac{1}{2}} \tag{5}$$

Finally, a compression, simulating basilar membrane compression, is applied to $e_i[m]$ to obtain the subband envelopes $s_i[m]$.

$$s_i[m] = f_{\text{comp}}(e_i[m]) = (e_i[m])^{0.3} \tag{6}$$

Once we have the subband envelopes, we calculate the statistics for the envelope mean, variance, skewness, kurtosis, cross correlation, the envelope modulation power, between subband (C1) modulation correlation and within subband (C2) modulation correlation. The variance of each subband, which is equivalent to the subband power, is also saved.

The residual of segment $x_{im}[n]$ is derived by dividing the segment by the envelope value.

$$r_{im}[n] = x_{im}[n]/e_i[m] \tag{7}$$

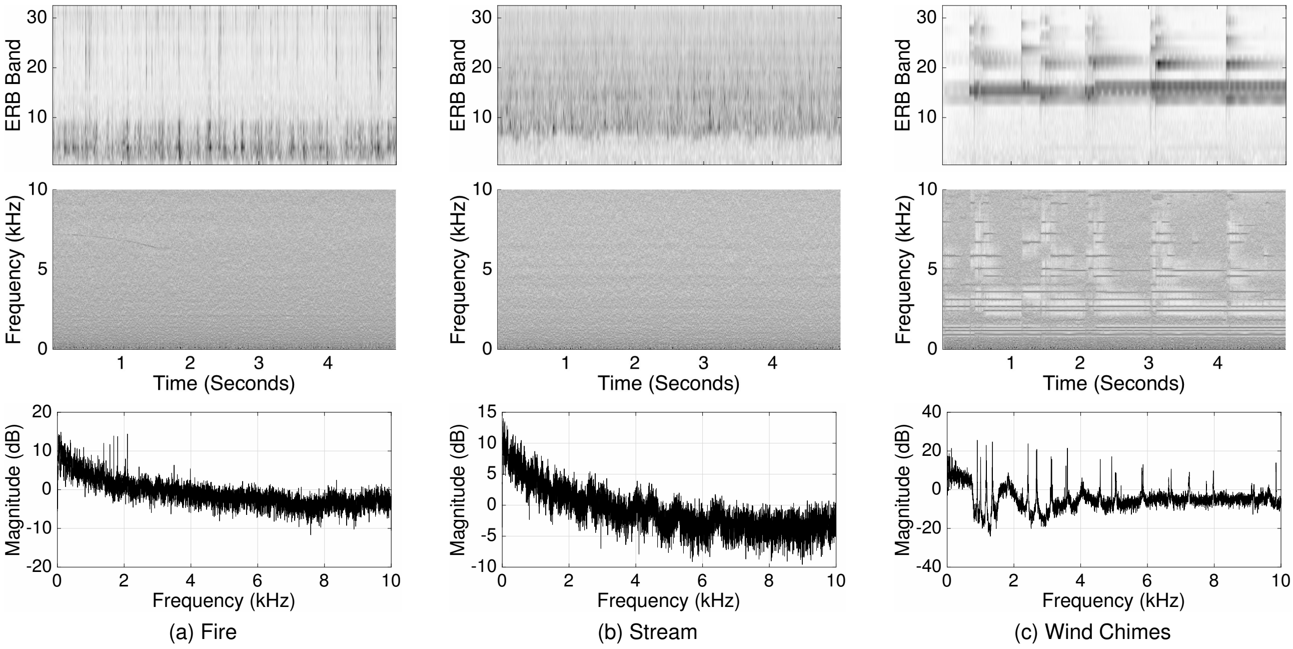
The segment residuals are merged to obtain the subband residual $r_i[n]$ and the subband residuals are summed to obtain the signal residual $r[n]$.

$$r_i[n] = \sum_m w[n + mR]r_{im}[n + mR] \tag{8}$$

$$r[n] = \sum_i h_i[n] * r_i[n] \tag{9}$$

While this decomposition process differs from McDermott and Simoncelli [16], the resulting envelope $s_i[m]$ is very similar. The envelope statistics imposing algorithm from McDermott and Simoncelli can be applied with little modification. The advantage of this formulation is that all the residuals are aggregated into one signal $r[n]$.

[1] This follows McDermott and Simoncelli's terminology. The term "modulation" is used to describe the frequency components of the envelopes.



(a) Fire

(b) Stream

(c) Wind Chimes

Figure 3: *Examples of sound texture decomposition and residual power spectral density. The first row shows the subband envelopes $s_i[m]$ of each signal. The second row shows the spectrogram of the residual $r[n]$. The third row is the power spectral density of the residual obtained using Welch's method. The y-axis of the envelope plot and the residual plot is different. (ERB scale vs. linear scale)*

3. RESIDUAL ANALYSIS

The envelopes and residuals from the sound texture decomposition can be viewed to have a carrier-modulation relation where the subband envelopes $s_i[m]$ are the amplitude modulations and the residual signal $r[n]$ is a temporally stable carrier signal. The subband envelopes, the spectrogram of the residual signal and the power spectral density (PSD) of the residual signal for example sound textures are shown in Figure 3.

The residual of crackling fire and flowing water is very close to pink (1/f) noise [21]. This is the result of normalizing the subband power over an ERB scale. Replacing the residual with pink noise for synthesis works well.

However, for a tonal sound like wind chimes (Figure 3c), the power spectral density is spiky due to the tonal components. Replacing the residual with pink noise would diffuse the tonal components, exciting the whole subband instead of focusing the signal power on a narrow band.

Inspecting the spectrogram of the residual in Figure 3c, the subband residuals do not look temporally stable, contrary to our assumption of carrier stability. Comparing the carrier spectrogram to the subband envelopes, we see that the subband envelopes have a small value where the tonals are missing. Thus, replacing the residual in Figure 3c with a temporally stable residual would not change the perceived output when merged with the subband envelopes.

Welch's method was used to estimate the power spectral density of the residual signal. We found that the shape of the tonal components is well captured when the averaging period is longer than 0.5 seconds. For the examples in this paper, an averaging period of 1 second was used at a sampling rate of 20 kHz.

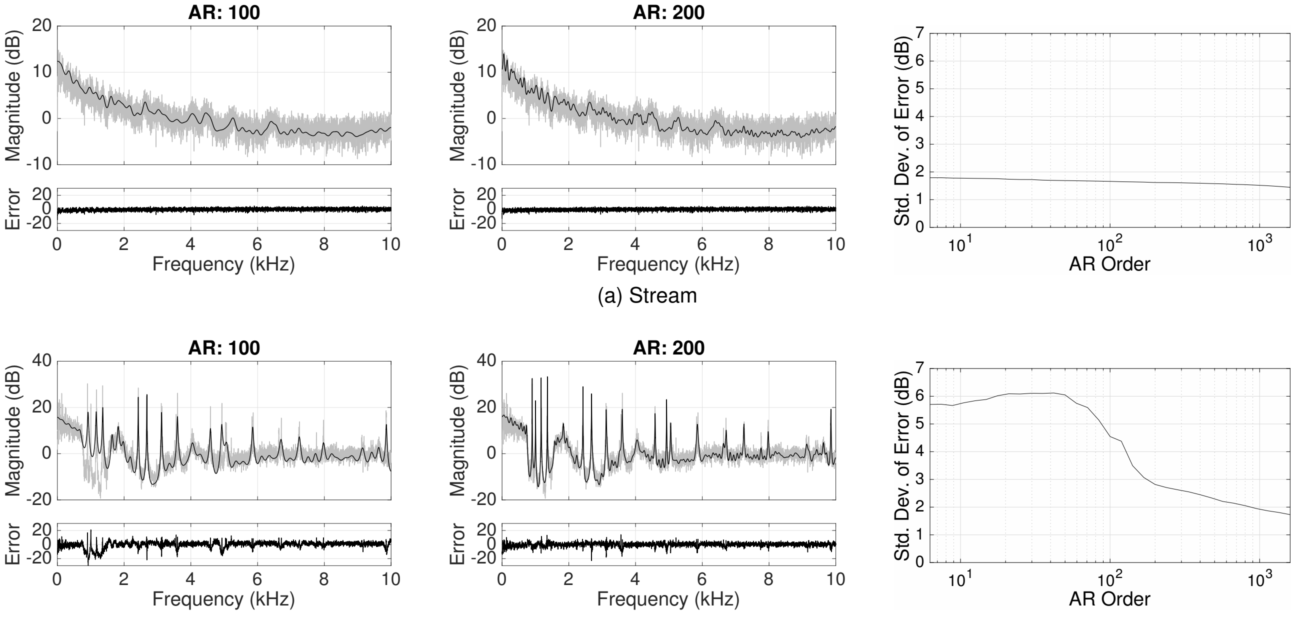
4. RESIDUAL MODELING

We can impose the power spectral density directly onto the residuals during the synthesis process to improve the results. Moreover, this will allow synthesis of sounds with tonal components. However, the amount of data for directly imposing the PSD is very large. For our example, 1 second at 20 kHz results in 10001 samples for the PSD. Much of the data is noise, we only need the contour of the PSD. One method of reducing the amount of data needed is by modeling the audio using high order auto-regressive (AR) modeling. High order AR modeling has been used for gap-filling and spectral modeling [22, 14]. The advantage of this approach is that we get high quality results without handling sinusoid components and noise components separately. A similar approach to tonal noise modeling has been covered by Polotti and Evangelista [23].

For non-tonal sounds, a good approximation can be obtained using low order AR models. However, for tonal sounds, it is important to model the tonal components well, especially the peak sharpness. If a tonal peak is modeled too broadly, that tonal component will sound diffused.

In Figure 4a, the residual is modeled evenly at both AR orders 100 and 200. In Figure 4b, the tonal components around 1kHz are not modeled well at an AR order of 100. Increasing the AR order to 200 improves the results.

To find a reasonable AR order, we plot the standard deviation of the magnitude errors against the AR order. For the stream example, there is little improvement with higher orders. For the wind chime example, we see a noticeable improvement between AR order 100 and 200. Examining more examples, an AR order of 200 was sufficient to model the tonal examples used for this paper.



(a) Stream

(b) Windchimes

Figure 4: *Residual AR modeling. In the first two columns, the frequency response of an AR model (black) is overlaid on the residual PSD (gray). Below each power spectrum, the error between the actual response and the AR approximation is plotted. In the third column, the standard deviation of the modeling errors is plotted for different AR orders. There is little improvement when increasing the AR order for the stream example. However, for the windchime example, we see a noticeable improvement between AR order 100 and 200.*

5. SOUND TEXTURE RESYNTHESIS

In this section, the process of extracting statistics and features from a source sound texture is covered with a detailed explanation of how the extracted statistics are used for resynthesis. The analysis and synthesis process is illustrated in Figure 5.

5.1. Extracting Sound Texture Statistics

After separating the source into subbands, the variance of each subband is saved. The subband variance is equivalent to the power of each subband signal. The human auditory system has acute sensitivity to the power in each subband, thus imposing the subband power correctly is important. The subband variance is used to “equalize” the subbands when resynthesizing.

Each subband is decomposed into its envelope and residual components as formulated in §3. The subband envelopes are then used to extract a subset of the statistics described in McDermott and Simoncelli [16]. We include modulation statistics in envelope statistics since the modulation statistics are all derived from the subband envelopes. One noticeable difference is that the envelopes are not windowed, windowing is inherently applied in the decomposition step. A detailed description of the statistics used is provided in §9.

The subband residuals are merged back into a full-band single channel residual signal as explained in equation (9). The AR coefficients are estimated from the residual signal (§4) and the AR coefficients are saved for use as an all-pole filter to synthesize a new residual signal.

5.2. Synthesizing from Sound Texture Statistics

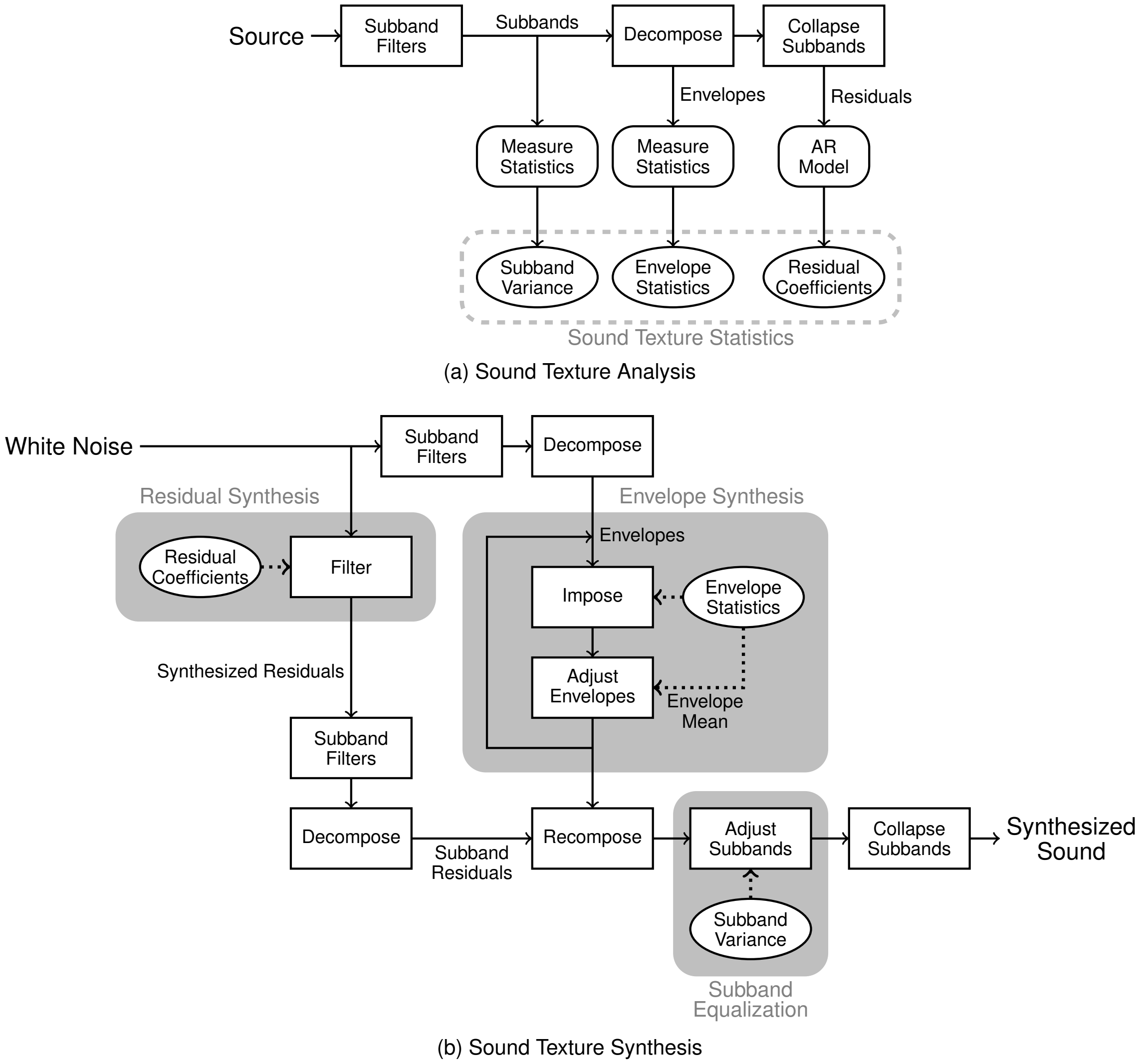
For resynthesis, starting with white noise, the envelopes and residuals are synthesized in parallel using the statistics from the analysis process. The two are then merged into subbands which are then equalized using the subband variances. The equalized subbands are then summed to form the final output signal.

5.2.1. Envelope Synthesis

After decomposing the subbands from the white noise signal into envelope and residual components, the residuals are discarded and only the envelopes are used for this step. The statistics imposing method was adapted from McDermott and Simoncelli [16]. For the target envelope statistics T_{env} extracted from the source sound texture and the current envelope statistics S_{env}, the L2 norm of $T_{env} - S_{env}$ is minimized using conjugate gradient descent.

Because the envelope mean is not normalized, it is not imposed in the gradient descent (See statistics formulas in §9). Instead, the envelope mean is imposed separately by adjusting the envelope means afterwards. It is worth noting that the uncompressed envelope $e_i[m]$, defined in equation (5), is proportionate to the power of the windowed segment $x_im[n]$ and thus the envelope mean is closely related to the subband power. However, because the synthesized residuals may not be spectrally flat, the subband variances are enforced after composing the synthesized residuals and envelopes to ensure the subband powers are correctly equalized.

This process is iterated until the difference between the target statistics and the current statistics is below a certain threshold or



(a) Sound Texture Analysis

(b) Sound Texture Synthesis

Figure 5: *Schematic of the sound texture analysis and synthesis procedure. The subband variance, envelope statistics and residual coefficients are measured and saved, then used for residual synthesis, envelope synthesis and subband equalization.*

the number of iterations pass a set limit. The process is not guaranteed to converge.

5.2.2. Residual Synthesis

The residual synthesis is straight forward. The input white noise signal is filtered with an all-pole filter composed of the AR coefficients from the analysis process. The synthesized residual is then decomposed into subband residuals and envelope components. The envelope components are discarded and the subband residuals are used for merging with the synthesized envelopes into subbands.

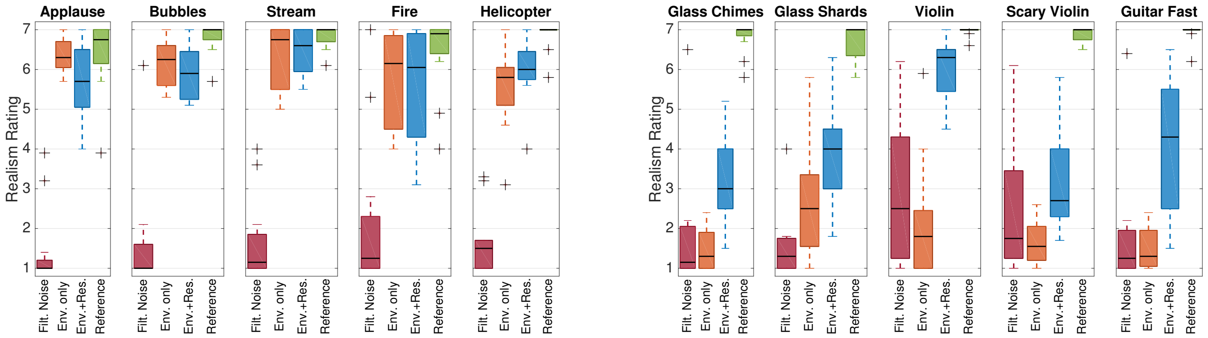
5.2.3. Equalization and Subband Rendering

Before merging the subbands, we adjust the variance of each subband. The equalization has a noticeable effect on the perception of the sound. In the recomposing step and collapse subband step the window $w[n]$ and the subband filters $h_i[n]$ are applied as synthesis windows and filters.

6. RESULTS

To test the effectiveness and validity of our model, we ran a user test where the participants were asked to rate the realism of resynthesized sound textures on a continuous scale from 1 to 7 with 1 being highly unrealistic and 7 being highly realistic.[2] Twelve subjects participated, 9 male, 3 female with a median age of 35. The participants were presented with the reference audio clip from which the statistics were extracted, along with 1 sample audio clip and 3 resynthesized audio clips, presented in random order. All audio clips were 4 seconds long.

[2] Sound samples used for the user tests are provided at `https://ccrma.stanford.edu/~hskim08/soundtextures/residual.html`.



(a) Non-tonal Sound Textures

(b) Tonal Sound Textures

Figure 6: *User test realism ratings. The median for each sample is shown with a thick black line in within the box. The box covers the first to third quartile (25% to 75%). The whiskers cover about 99%. The + symbols represent the outliers. The realism scale provided to the users is 1-Highly unrealistic, 2-Unrealistic, 4-Acceptible, 6-Realistic, 7-Highly Realistic.*

The sample clip was taken from a different part of the same audio file as the reference audio. The three methods of resynthesis were 1) white noise filtered to match the PSD of the source audio, 2) white noise with only the envelope synthesized, and 3) white noise with both envelope synthesis and residual synthesis. The first method simulates residual only resynthesis, while the second case only uses the envelope statistics for resynthesis.

For the non-tonal sound textures, both the envelope only case and the case with both envelope and residual synthesis were perceived to have high realism. Filtered noise was perceived to contain low realism. For these examples, it seems most of the perceived information is in the envelopes and thus synthesizing the envelopes was sufficient to create realistic samples.

For sound textures with tonal components, the effect of the residual synthesis becomes visible. For violin sounds, filtered noise scaled higher realism than the envelope synthesis case, implying that the spectral information was more dominant in the perception of those sounds. In all cases, using both envelopes and residual for synthesis was perceived to be more realistic those that used only one.

Two examples worth noting are that of fire and violin. Despite having no tonal component, the fully resynthesized sample of fire was perceived to have lower realism than the other non-tonal examples. The fully resynthesized sample of violin on the other hand was perceived to have very high realism compared to other tonal examples. We believe this is due to the limitations of the temporal subband shaping modeled with the within subband (C2) modulation correlation. The crackling in fire as well as the attacks of the tonal examples have a noticeable temporal effect. However, we have found that the effects of enforcing the C2 correlation seemed to be limited. When the C2 correlation is easy to match, as in the violin example, we see that our method creates very compelling results.

7. CONCLUSIONS

We presented a method of decomposing a sound texture into its envelope and residual components. Examining the residuals for different examples, it was observed that non-tonal sound textures had residuals with power spectral densities close to pink noise, while that was not the case for tonal sound textures. Thus, the need for residual modeling. Applying high order auto-regression modeling to the residual, it was possible to reduce the data needed by a magnitude of two with little perceived differences. We presented a system for extracting the statistics from a source sound texture and the system for using the statistics to resynthesize new examples. The importance of both the residuals and envelopes was verified by a user test. For non-tonal sounds, a good envelope model was sufficient to synthesize realistic sounds. Adding the residual modeling did not affect the realism. However, for tonal sounds, modeling both the residuals and envelopes gave more realistic results than modeling just the residuals or the envelopes.

Taking a higher point of view, our approach fills the gap between filtered noise and the sound texture analysis presented by McDermott and Simoncelli [16]. Filtered noise captures the short term statistics in the form of power spectral distribution, including tonal components. Meanwhile summary statistics capture the modulations on the order of seconds.[3] Revisiting Saint-Arnaud's comparison of speech, music, noise, and sound textures in Figure 1, our model provides an explanation for the intuition behind the relation between potential information and time. The spectral distribution for noise can be estimated in a few milliseconds, while the subband modulations can be estimated on the order of seconds. The structure of speech and music is defined over a time span greater than that of seconds, usually minutes or longer.

The original objective of the study was to improve the sound texture model of McDermott and Simoncelli to cover tonal sound textures such as wind chimes. Over the course of time, we found that the model could be applied to constant pitch sounds such as a single note on a violin or guitar. We could model the textural aspect of the instrument sound such as fast bowing or tremolo picking. This suggests that the analysis of modulations could be applied to instrument modeling to add textural timbres.

While AR modeling was used to reduce the amount of data needed to synthesize the residuals, a sines+noise like approach could futher reduce the data. We were able to model the residuals of non-tonal sound textures sufficiently using AR orders of 10. By separately modeling the tonal peaks of the PSD, then using AR modeling only on the remaining residuals, it seems possible to reduce the amount of data by another order.

[3] The lowest modulation band used is 0.5Hz. See §9.

During the user test, the limited effectiveness of within subband (C2) modulation correlation enforcement for temporal modeling was observed. Improvements in temporal modeling seems to be an important factor in increasing the realism of the proposed sound texture synthesis method.

For this study, we limited the tonal sound textures to those that do not have variable pitch trajectories. This excludes most cases of vocalizations including bird songs, babies crying and speech. These sounds may require a completely different approach as there may be tonal components that move between subbands which may be challenging to model. Once more a sines+noise decomposition may prove to be useful for such cases, where the tonal components are modeled separately and the noise component, din, could be modeled by our proposed method.

Finally, there is a lack of evaluation metrics for sound textures. Evaluating the samples with PQevalAudio [24], all samples scored a very low objective difference grade, -3.5 or less, on a scale of 0 to -4 where 0 is imperceivable and -4 is very annoying. This seems to be caused by the fact that PQevalAudio compares the audio on a frame to frame basis meaning that it compares short term statistics whereas the synthesis for sound texture enforces long term statistics and as such the short term statistics can be very different. This is likely the case for other perceptual audio evaluation metrics. The short term measurements for sound textures may vary, yet the perception of the sounds are similar [25], suggesting that a different metric would be needed to programatically evaluate the perceived quality of resynthesized sound textures. Validating sound texture models with improved analysis/synthesis results should help make better perceptual evaluation metrics.

8. REFERENCES

[1] Nicolas Saint-Arnaud, "Classification of Sound Textures," M.S. thesis, Massachusetts Institute of Technology, 1995.

[2] Gerda Strobl, Gerhard Eckel, and Davide Rocchesso, "Sound Texture Modeling: A Survey," pp. 61–65, 2006.

[3] Diemo Schwarz, "State of the Art in Sound Texture Synthesis," in *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx-11)*, Paris, France, September 2011.

[4] Nicolas Saint-Arnaud and Kris Popat, "Analysis and Synthesis of Sound Textures," in *Readings in Computational Auditory Scene Analysis*, 1995, pp. 125–131.

[5] Ziv Bar-Joseph, Ran El-Yaniv, Dani Lischinski, Michael Werman, and Shlomo Dubnov, "Granular Synthesis of Sound Textures using Statistical Learning," in *Proceedings of the International Computer Music Conference*, Beijing, China, 1999.

[6] Lie Lu, Liu Wenyin, and Hong-Ziang Zhang, "Audio Textures: Theory and Applications," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 156–167, Mar. 2004.

[7] Ananya Misra, Ge Wang, and Perry Cook, "Musical Tapestry: Re-composing Natural Sounds," *Journal of New Music Research*, vol. 36, no. 4, pp. 241–250, Dec. 2007.

[8] Gerda Strobl, *Parametric Sound Texture Generator*, Ph.D. thesis, Graz University of Technology, 2007.

[9] Martin Fröjd and Andrew Horner, "Sound Texture Synthesis Using an Overlap–Add/Granular Synthesis Approach," *Journal of the Audio Engineering Society*, vol. 57, no. 1/2, pp. 29–37, 2009.

[10] Shlomo Dubnov, Naftali Tishby, and Dalia Cohen, "Polyspectra as Measures of Sound Texture and Timbre," *Journal of New Music Research*, vol. 26, no. 4, pp. 277–314, Dec. 1997.

[11] Shlomo Dubnov, Ziv Bar-Joseph, Ran El-Yaniv, Dani Lischinski, and Michael Werman, "Synthesizing Sound Textures Through Wavelet Tree Learning," *Computer Graphics and Applications, IEEE*, vol. 22, no. 4, pp. 38–48, 2002.

[12] Doug Van Nort, Jonas Braasch, and Pauline Oliveros, "Sound Texture Analysis Based on a Dynamical Systems Model and Empirical Mode Decomposition," in *Audio Engineering Society Convention 129*, Nov 2010.

[13] Joan Bruna and Stéphane Mallat, "Audio Texture Synthesis with Scattering Moments," *arXiv.org*, Nov. 2013.

[14] Marios Athineos and Daniel PW Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 5, pp. V–648.

[15] Xinglei Zhu and Lonce Wyse, "Sound Texture Modeling and Time-Frequency LPC," in *Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx-04)*, Naples, Italy, October 2004.

[16] Josh H McDermott and Eero P Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, Sept. 2011.

[17] Wei-Hsiang Liao, Axel Roebel, and Alvin WY Su, "On the Modeling of Sound Textures Based on the STFT Representation," in *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, September 2013.

[18] Xavier Serra and Julius Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[19] Brian R Glasberg and Brian CJ Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[20] Parishwad P Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, 1993.

[21] Bruce J West and Michael F Shlesinger, "On the Ubiquity of 1/f Noise," *International Journal of Modern Physics B*, vol. 03, no. 06, pp. 795–819, June 1989.

[22] Paulo AA Esquef and Luiz WP Biscainho, "An Efficient Model-based Multirate Method for Reconstruction of Audio Signals Across Long Gaps," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1391–1400, 2006.

[23] Pietro Polotti and Gianpaolo Evangelista, "Fractal additive synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 105–115, March 2007.

[24] Peter Kabal, "An Examination and Interpretation of ITU-R BS. 1387: Perceptual Evaluation of Audio Quality," *TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University*, pp. 1–89, 2002.

[25] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli, "Summary Statistics in Auditory Perception," *Nature Neuroscience*, vol. 16, no. 4, pp. 493–498, Feb. 2013.

9. APPENDIX: ENVELOPE STATISTICS

The envelope statistics are adapted from McDermott and Simoncelli [16], with minor changes to accommodate the differences in how the subband envelopes $s_i[m]$ were derived. The most noticeable difference is the replacement of the window function $w(t)$ with $1/N$. We formulate the statistics here for completeness.

The envelope statistics can be categorized into subband envelope statistics and envelope modulation statistics. The subband statistics are directly measured from the subband envelopes, whereas the modulation statistics are measured after the subbands are filtered into modulation bands through a constant Q filter bank $\bar{f}_u[m]$ for the modulation power and an octave spaced filter bank $f_u[m]$ for the C1 and C2 correlations.

9.1. Subband Envelope Statistics

We start by defining the envelope moments. Defining the moments help simplify the definitions of the marginals. Precalculating the moments can reduce computation times. For the i-th subband envelope $s_i[m]$, the envelope moments are defined as,

$$\mathrm{m}_1[i] = \mu_i = \frac{1}{N}\sum_{m=1}^{N} s_i[m]$$

$$\mathrm{m}_X[i] = \frac{1}{N}\sum_{m=1}^{N}\{s_i[m]-\mu_i\}^X, \qquad X > 1$$

The standard deviation σ_i is also useful to precalculate.

$$\sigma_i = \sqrt{\mathrm{m}_2[i]}$$

9.1.1. Envelope Marginals

The envelope marginals, except for the envelope mean $\mathrm{M1}_i$, are normalized. This makes the statistics independent from any scaling factors. This is also important when imposing the statistics using optimization. Because the envelope mean is not normalized and tends to have smaller values than all other statistics used, it needs to be enforced separately after the optimization. The envelope marginals help shape the general distribution of the envelopes.

$$\mathbf{M1_i} = \mathrm{m}_1[i] = \mu_i$$

$$\mathbf{M2_i} = \frac{\mathrm{m}_2[i]}{(\mathrm{m}_1[i])^2} = \{\frac{\sigma_i}{\mu_i}\}^2$$

$$\mathbf{M3_i} = \frac{\mathrm{m}_3[i]}{(\mathrm{m}_2[i])^{3/2}} = \frac{1}{N}\frac{\sum_{m=1}^{N}(s_i[m]-\mu_i)^3}{\sigma_i^3}$$

$$\mathbf{M4_i} = \frac{\mathrm{m}_4[i]}{(\mathrm{m}_2[i])^2} = \frac{1}{N}\frac{\sum_{m=1}^{N}(s_i[m]-\mu_i)^4}{\sigma_i^4}$$

9.1.2. Envelope Cross-band Correlation

This is the correlation coefficient of two subband envelopes $s_i[m]$ and $s_j[m]$.

$$\mathrm{C_{ij}} = \frac{1}{N}\sum_{m=1}^{N}\frac{(s_i[m]-\mu_i)(s_j[m]-\mu_j)}{\sigma_i\sigma_j}$$

The envelope cross-band correlation helps enforce the comodulation of the subbands.

9.2. Envelope Modulation Statistics

Each subband envelope is further decomposed into its modulation bands through another filter bank. The modulation bands cover frequencies from 0.5Hz to $f_{env} = 400$Hz. Two different filter banks are used for the modulation power and the C1/C2 modulation correlations. The modulation power is calculated using a constant Q filter bank $\bar{f}_u[m]$.

$$\bar{b}_{i,u}[m] = \bar{f}_u[m] * s_i[m]$$

The C1/C2 modulation correlations are calculated using an octave band filter bank $f_u[m]$. An octave band is chosen because of the formulation of the C2 correlation.

$$b_{i,u}[m] = f_u[m] * s_i[m]$$

9.2.1. Modulation Power

The modulation power $\mathrm{M}_{i,u}$ is the root-mean-square of the modulation band normalized by the variance of the whole subband σ_i^2. The modulation power can be viewed as the distribution of the subband power within the modulation bands.

$$\mathrm{M_{i,u}} = \frac{1}{N}\frac{\sum_{m=1}^{N}(\bar{b}_{i,u}[m])^2}{\sigma_i^2}$$

9.2.2. Between Band Modulation (C1) Correlation

The C1 correlation is the correlation coefficient of two subband modulations $b_{i,u}[m]$ and $b_{j,u}[m]$ where i and j are the subband numbers and u is the modulation band number.

$$\mathbf{C1_{ij,u}} = \frac{1}{N}\sum_{m=1}^{N}\frac{b_{i,u}[m]b_{j,u}[m]}{\sigma_{i,u}\sigma_{j,u}}$$

where,

$$\sigma_{i,u} = \sqrt{\frac{1}{N}\sum_{m=1}^{N} b_{i,u}[m]}$$

The C1 correlation helps enforce the comodulation of subbands within the same modulation band.

9.2.3. Within Band Modulation (C2) Correlation

The C2 correlation enforces the temporal shape of a subband by imposing phase of modulation bands within a subband. To compare the phase of adjacent subbands, the modulation bands are transformed to its analytic signal $a_{i,u}$.

$$a_{i,u}[m] = b_{i,u}[m] + j\mathcal{H}\{b_{i,u}[m]\}$$

Next, the lower octave signal is expanded an octave by squaring the values, then normalized.

$$d_{i,u}[m] = \frac{(a_{i,u}[m])^2}{\|a_{i,u}[m]\|}$$

The correlation coefficient of the two bands is calculated for the C2 correlation.

$$\mathbf{C2_{i,uv}} = \frac{1}{N}\sum_{m=1}^{N}\frac{d_{i,v}^{*}[m]a_{i,u}[m]}{\sigma_{i,u}\sigma_{i,v}}$$