Analysis and synthesis of environmental sounds based on auditory principles

Jan Brümmerstedt[1], Richard McWalter[1], Torsten Dau[1]

[1] CAHR, Technical University of Denmark, 2900 Kongens Lyngby, Denmark, Email: jan.bruemmerstedt@gmx.de

Introduction

The analysis via synthesis approach has been used to examine auditory perception. The general idea is to capture a set of statistics from an input signal at the analysis stage, and then synthesise a new instance based on those statistics. If the perceptually relevant statistics are measured from a biologically inspired auditory model, the original and the synthetic sounds should be perceptually similar. Thus, the analysis via synthesis approach provides a strong test of a perceptual model. [3] have applied such an auditory model to the synthesis of textures - temporally homogenous sounds, such as rain or birds chirping. They showed that synthetic textures generated from a biologically inspired model were preferred over those generated from non-biological models. This synthesis method is applicable to sound textures, but does not account for the temporal fluctuations common in many environmental sounds. In contrast to this long-term model, we propose an extension to account for the dynamic and unique features of environmental sounds. This was achieved by introducing higher temporal resolution in the modulation domain, yielding the multi-resolution analysis-synthesis model. The individual contributions of the various statistics in the spectral domain as well as the modulation domain were examined. Lastly, the effect of decoupling the modulation power statistics from the spectral domain was discussed.

Model

The proposed algorithm consists of an auditory model serving as front-end, embedded in an analysis-synthesis framework.

Auditory Model

The auditory model consists of three processing stages: Frequency selective (peripheral) filtering, envelope extraction and compression, and finally, the modulation processing as proposed by [1]. The representation of the signal after the frequency selective filtering, envelope extraction and compression, is defined here as envelope domain, while the representation of the signal after the modulation processing is defined here as modulation domain. The statistics captured in these domains are defined as envelope statistics and modulation statistics, respectively. After the envelope extraction, the signal envelopes are downsampled from 20 kHz to 400 Hz, yielding a Nyquist frequency of 200 Hz.

Peripheral Filterbank

The peripheral filtering is implemented as a gammatone filter bank of order $n = 4$ with a tuning parameter of $b = 1.0183$. A gammatone filter is defined as:

$$\gamma[n] = an^{v-1}e^{-\lambda n}e^{2\pi i f_c n}, \tag{1}$$

with center frequency f_c, amplitude a, and the damping factor $\lambda = 2\pi b\mathrm{ERB}(f_c)$. The equivalent rectangular bandwidth (ERB) of a human auditory filter was estimated by [2] as $\mathrm{ERB}(f_c) = 24.7 + 0.108f_c$. The tuning parameter b sets the bandwidth of the filter in relation to the ERB. The order $n = 4$ and tuning parameter $b = 1.0183$ are derived from a notched-noise masking experiment. The gammatone design is chosen to reflect the frequency selectivity of the peripheral auditory system. The output are 34 channels, or sub-bands, with center frequencies ranging from 50 Hz to ≈ 8 kHz.

Envelope Extraction and compressive Non-Linearity

[1] showed that the TMTF depends on the carrier wave form. This is due to the intrinsic modulations of the carrier. In analogy to the power spectrum model, the intrinsic envelope fluctuations of the carrier mask the imposed sinusoidal modulation. In this model, the envelope is extracted by applying the Hilbert transform, given in equation (2).

$$S_H(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{s(\tau)}{\tau - t}d\tau \tag{2}$$

By applying equation (2), one can construct the corresponding analytic signal, given by $s_a(t) = s(t) - iS_H(t)$. The Hilbert envelope can then be extracted by taking the modulus of the analytic signal, according to equation (3). The modulation spectrum is then the power spectrum of the Hilbert envelope.

$$h_{env}(t) = |s_a(t)| = \sqrt{s^2(t) + S_H^2(t)} \tag{3}$$

This leads to a modulation masking pattern, which in turn leads to the formulation of the modulation filter bank. By applying this modulation filter bank, one is able to predict the TMTF for various carriers.

After the envelope extraction a compressive non-linearity is applied in the form of a static exponential.

Modulation Filterbank

The last stage in the model is the modulation filter bank as proposed by [1]. It is implemented as a bank of 8 FIR bandpass filters with Kaiser window and constant relative bandwidth. The center frequencies are octave spaced and range from $f_c = 1.25$ Hz to $f_c = 160$ Hz. The modulation bands with their respective center frequencies f_c and window durations $L(w_n)$ are listed in table 1. The

centre frequency of 1.25 Hz and the subsequent octave spacing were chosen due to the divisibility of the down-sampled envelope frequency (400Hz) to round numbers. By applying time windows of different durations, the model is effectively implementing different time constants τ_n.

Tabelle 1: Modulation filter channels

Band	1	2	3	4	5	6	7	8
f_c [Hz]	1.25	2.5	5	10	20	40	80	160
$L(w_n)$	320	160	80	40	20	10	5	2
τ_n [ms]	800	400	200	100	50	25	12.5	5

Analysis-Synthesis Framework

The analysis-synthesis framework, as shown in figure 1, allows to synthesise new instances of a sound, based on its statistical properties. In the analysis part, the target signal is processed by the auditory model described above, and then a set of statistics are captured at the envelope and modulation domain, as described below. These statistics are transferred to the synthesis part, where a random noise sample is decomposed based on the same auditory model, and the statistics from the target signal are imposed. The imposition of the statistics is implemented as a gradient projection. Since this is an approximative process, it is iterated until the statistics of the synthetic signal are sufficiently close to the target signal. This is estimated by a statistical signal-to-noise ratio (SNR) for each statistic as defined in equation (4). As convergence criterion, it has been set, that all relevant statistics have to have a statistical SNR greater than 30 dB.

$$\mathrm{SNR}_{\mathrm{stat}} = 10\log_{10}\frac{\sum_k(\sum_n)(\sum_l)X^2_{\mathrm{stat}}}{\sum_k(\sum_n)(\sum_l)(X_{\mathrm{stat}}-Y_{\mathrm{stat}})^2}, \quad (4)$$

with sub-bands $k \in [1...34]$, and (in the case of the modulation statistics) modulation bands $n \in [1...8]$, as well as time windows $l \in [1...l_N]$.

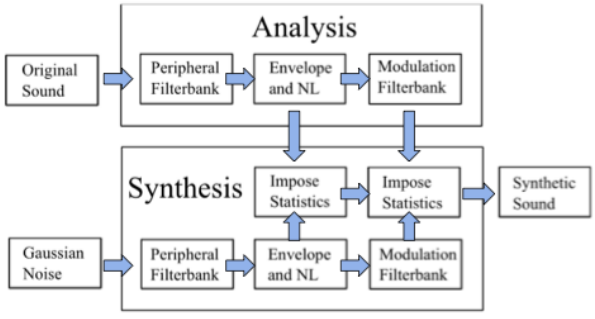

Abbildung 1: Analysis-synthesis framework, including the decomposition of the input signals by a gammatone filter bank, envelope extraction, a compressive non-linearity and a modulation filter bank. Input to the analysis section is the original signal, input to the synthesis section a random noise sample. Output is a synthetic signal with the closely matched target statistics.

Statistical Set

In this framework, statistics are captured respectively imposed on two levels: the envelope domain and the modulation domain.

Envelope domain

The first four statistical moments at the output of each of the 34 filters are computed, namely:

1. Mean, $M1_k = \mu_k = \sum_t w(t)s_k(t)$,
2. Coefficient of Variance, $M2_k = \frac{\sigma_k^2}{\mu_k^2} = \frac{\sum_t w(t)(s_k(t)-\mu_k)^2}{\mu_k^2}$,
3. Skewness, $M3_k = \frac{\sum_t w(t)(s_k(t)-\mu_k)^3}{\delta_k^3}$,
4. Kurtosis, $M4_k = \frac{\sum_t w(t)(s_k(t)-\mu_k)^4}{\delta_k^4}$,

for the k^{th} envelope sub-band $s_k(t)$, and the windowing function $w(t)$, with the constrain that $\sum_t w(t) = 1$. Note that each marginal is normalised to its respective filter. In order to get a dimensionless value, the variance has been normalised with $1/\mu^2$, thus giving the coefficient of variance. Note further, that the higher order moments, skewness and kurtosis, are not imposed in the default version of the proposed multi-resolution model.

Additionally, across-channel correlations were calculated as given by equation (5),

$$C_{jk} = \sum_t \frac{w(t)(s_j(t)-\mu_j)(s_k(t)-\mu_k)}{\delta_j\delta_k}, j,k \in [1...34], \quad (5)$$

such that $(k-j) \in [1,2,3,5,8,11,16,21]$.

Modulation domain

The modulation power is computed as:

$$M_{k,n,l} = \frac{\sum_t w_{n,l}w(t)b_{k,n}(t)^2}{\delta_k^2}, \quad (6)$$

with $k \in [1...34], n \in [1...8], l \in [1...n_w]$ and $w_{n,l}$ designating the l^{th} window in the n^{th} modulation channel $b_{k,n}$. The total number of windows n_w depends on the modulation channel and the signal duration. Note, how the modulation power is normalised with regards to the global sub-band variance δ_k^2. Implementations of a local sub-band variance $\delta_k^2 w_{n,l}(t)$ as normalisation factor failed in producing sufficient synthesis results. In order to capture only meaningful statistics, the modulation statistics have been limited to modulation channels with a centre frequency $f_{c,k,n}$ smaller than a quarter of the sub-band envelope centre frequency $f_{c,k}$, $f_{c,k,n} < f_{c,k}/4$ [1]. This is reasonable, as it does not make much sense to capture the 160 Hz modulation of a sub-band channel centred around 50 Hz.

Method

Two subjective listening experiments were conducted. Experiment 1 compared the synthesis quality of the proposed multi-resolution framework against that of the original long-term model by [3]. Experiment 2 examined the influence of the gradual statistics by comparing implementations of varying complexity of the proposed model.

Both experiments were conducted as Multi-Stimulus with Hidden Reference and Anchor (MUSHRA) test [4], where subjects were seated in a double-walled sound attenuated booth and presented with different stimuli. These stimuli they had to rate according to their realism on a scale from 0 to 100. All subjects were self reported normal hearing, and between 21 and 33 years old. Both tests were conducted over 24 different environmental sounds, comprised of 17 sound textures and 7 non-texture sounds. The reference condition was always the original sound, and the low anchor the long-term synthesis based on the envelope power only.

The conditions for experiment 1 were:

- Original (Reference)
- Multi-Resolution Model
- Long-term Model
- Envelope Power (Low Anchor)

The conditions for experiment 2 were:

- Original (Reference)
- Multi-Resolution Model without correlations
- Multi-Resolution Model
- Multi-Resolution Model incl. skewness and kurtosis
- Envelope Power (Low Anchor)

Results

Experiment 1

Figure 2 shows the results for the test comparing the proposed multi-resolution model (MR) against the long-term model (LT) from [3], grouped over all subjects and sound files. A two-factor repeated measures analysis of variance (ANOVA) showed that there is a significant effect of the condition on the quality rating ($F(1,8) = 37.054, p < 0.001$). Note, that the ANOVA was performed only over the two synthesis models, excluding the hidden reference and the low anchor. The proposed MR model shows a significant improvement of 11.273 points (distance d_m between the means) over the LT model even averaged over all sound files.

Figure 3 shows the results for experiment 1 grouped into sound textures (left) and non texture sounds (right). When analysed over these subgroups, the two models show a smaller difference when considering the sound textures ($d_m = 7.500, p = 0.016$), and a much larger improvement of the MR model when considering the non-texture sounds ($d_m = 22.593, p < 0.001$). This is in line with the hypothesis, as the LT model was developed for sound textures, while the MR model was developed to specifically address non-texture environmental sounds. It is worth mentioning though, that while the difference is statistically significant, the data in general exhibit a large spread.

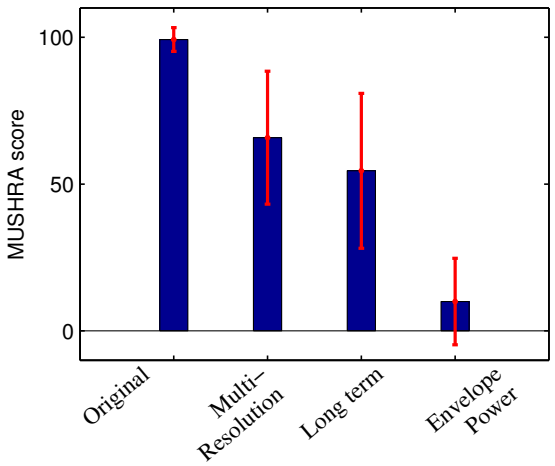

Abbildung 2: Results of the experiment on the comparison of the multi-resolution and the long-term model, showing mean and standard deviation over subjects and sound files. Conditions from left to right: (1) Original sound (hidden reference), (2) multi-resolution model, (3) long-term model from [3], (4) envelope power shaped noise (low-anchor)

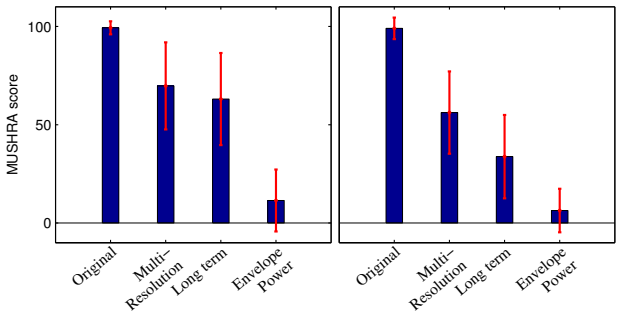

Abbildung 3: Results of the experiment on the comparison of the multi-resolution and the long-term model, showing mean and standard deviation over subjects and sound files, for the subgroups sound textures (left) and non-texture sounds (right). Conditions from left to right: (1) Original sound (hidden reference), (2) multi-resolution model, (3) long-term model from [3], (4) envelope power shaped noise (low-anchor)

Experiment 2

Figure 4 shows the results of the experiment on gradual statistics grouped over all subjects and all sound files. A two-factor repeated measures ANOVA showed that there is a significant effect of the condition on the quality rating ($F(2,12) = 17.840, p < 0.001$). Again, the ANOVA was performed only over the synthesis model conditions, excluding the hidden reference and the low anchor. In order to examine the differences between the different models, a post-hoc analysis via a multiple comparison t-test with Bonferroni correction has been performed. This showed statistical significance between conditions 2 and 3 ($p = 0.011$), and conditions 2 and 4 ($p = 0.009$), but not between conditions 3 and 4 ($p = 0.128$). This means, that omitting correlation statistics lead to minor, but significant decrease in synthesis quality. Adding the higher order moments, skewness and kurtosis, on the other hand, was not found to significantly improve the performance.

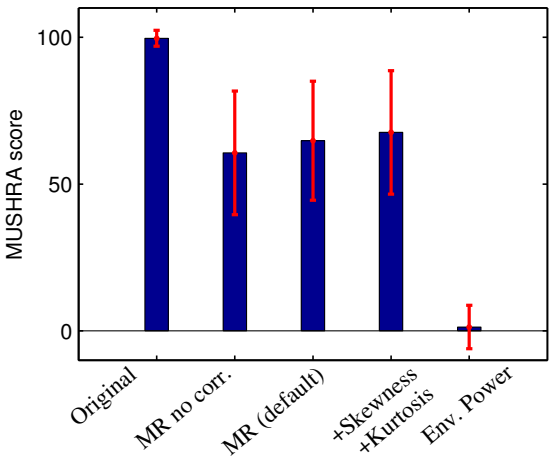

Abbildung 4: Results of the experiment on gradual statistics of the multi-resolution model, showing mean and standard deviation over subjects and sound files. Conditions from left to right: (1) Original sound (hidden reference), (2) multi-resolution model without correlations, (3) multi-resolution model, (4) multi-resolution model including skewness and kurtosis, (5) envelope power shaped noise (low-anchor)

In figure 5, the sound files are grouped into sound textures (left) and non-texture sounds (right). The latter class was expected to exhibit greater differences between the conditions, as it is the temporally more complex signal class. Again, the two-factor repeated measures ANOVA showed a significant difference between the models ($F(2,12) = 12.386, p < 0.001$ for the sound textures, $F(2,12) = 6.468, p < 0.012$ for the non-texture sounds). A post-hoc multiple comparison test with Bonferroni correction on the sound textures showed a significant difference between conditions 2 and 3 (distance between the means $d_m = 3.611, p = 0.046$), and conditions 2 and 4 (with similar $d_m = 3.151, p = 0.011$), but not between conditions 3 and 4 ($p = 0.237$). For the non-texture sounds the Bonferroni-corrected multiple comparison test showed only a significant difference between conditions 2 and 4 ($d_m = 7.619, p = 0.028$), but not between conditions 2 and 3 ($p = 0.209$), and conditions 3 and 4 ($p = 1$). The quality increase for the included correlations might be relatively minor, but for some sounds, even if few in number in this test set, they yield a perceivable difference.

Discussion

The multi-resolution model allows to generate synthetic sounds perceptually similar to their corresponding target signals. One issue, however, is the decoupling of the modulation domain statistics from the envelope domain statistics. A proper decoupling on a local scale could not be sufficiently implemented, therefore only a global decoupling was implemented. Without this local decoupling, the modulation domain statistics inform and shape the subband. Due to the increased temporal resolution in the modulation domain, this brings the proposed model conceptually closer to the noise vocoder by [5]. However, it seems that a total independence of the statistics is not

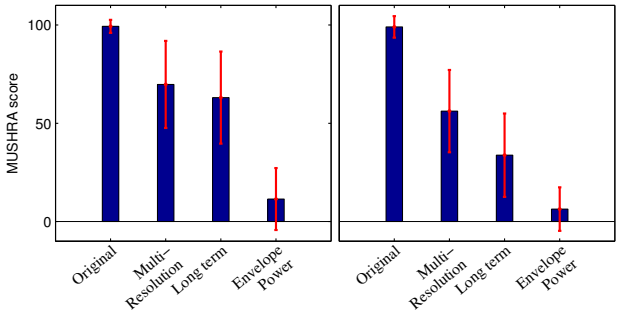

Abbildung 5: Results of the experiment on gradual statistics of the multi-resolution model, showing mean and standard deviation over subjects and sound files, for the subgroups sound textures (left) and non-texture sounds (right). Conditions from left to right: (1) Original sound (hidden reference), (2) multi-resolution model without correlations, (3) multi-resolution model, (4) multi-resolution model including skewness and kurtosis, (5) envelope power shaped noise (low-anchor)

achievable, as was also found by [6].

Conclusion

It could be shown that the multi-resolution processing scheme improves the perceived synthesis quality, especially for the temporally complex non-texture sounds. Further on, the influence of some of the statistics on the synthesis quality decreased. These were skewness, kurtosis and the across-channel correlations. Omission of the correlation statistics lead to minor, but significant decrease in synthesis quality. Since adding the higher order moments, skewness and kurtosis, on the other hand, was not found to significantly improve the performance, these were not included in the proposed model.

Literatur

[1] Dau, T., Kollmeier, B. and Kohlrausch, A.: Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers, JASA 102 (1997), 2892-2905

[2] Moore, B., Peters, R. and Glasberg, B.: Auditory filter shapes at low center frequencies, JASA 88 (1990), 132-140

[3] McDermott, J., Schemitsch, M. and Simoncelli, E.P, Summary statistics in auditory perception. Nature neuroscience 16 (2013), 493-498

[4] Recommendation ITU-R BS.1534-2: Method for the subjective assessment of intermediate quality level of audio systems, ITU 2014

[5] Shannon, R., Zeng, F., Kamath, V., Wygonski, J. and Ekelid, M.: Speech recognition with primarily temporal cues, Science 270 (1995), 303-304

[6] McDermott, J. and Simoncelli, E.P, Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis., Neuron 71 (2011), 926-940