Towards equalization of environmental sounds using auditory-based features

Jorge Garcia  
Music Technology Group  
Universitat Pompeu Fabra  
Barcelona, Spain  
neo_jgm@yahoo.es

Stefan Kersten  
Music Technology Group  
Universitat Pompeu Fabra  
Barcelona, Spain  
stefan.kersten@upf.edu

Jordi Janer  
Music Technology Group  
Universitat Pompeu Fabra  
Barcelona, Spain  
jordi.janer@upf.edu

ABSTRACT
In this paper we describe methods to assist soundscape design, sound production and processing for interactive environments, like games and simulations. Using auditory filter banks and sound texture synthesis, we develop algorithms that can be integrated with existing audio engines and can additionally support the development of dedicated high-level audio tools aimed at content authoring or transformations based on samples. The relationship between the auditory excitation patterns and the computation algorithm is explained within the context of footstep sounds. Moreover, methods for sound texture synthesis of water streams with artificial expansion of timbre space using auditory filtering techniques are presented.

Categories and Subject Descriptors
H.5.5: [Information Interfaces and Presentation]: Sound and Music Computing - Methodologies and techniques, Modeling, Signal analysis, synthesis and processing.

General Terms
Algorithms, Design.

Keywords
Analysis, auditory filtering, equalization, texture synthesis.

1. INTRODUCTION
Sound design is still a slow and manual process, where expert skills and access to large sound effects libraries, or complex synthesis algorithms are required. Simultaneously, we observed in recent years a proliferation of collaborative media repositories for images, video and sound. In this context, we aim to exploit sound material from the user-contributed database Freesound (http://www.freesound.org). A relevant characteristic of this repository is that sound samples for a given same concept (e.g. “soft rain”) can be created by different users, and thus differ in the actual recorded sound source and recording conditions.

In the design of interactive soundscapes [1], we can combine several samples for generating a sound concept, obtain a richer and non-repetitive sound that can run autonomously (e.g. by sequencing it with a graph model). To improve the realism of the synthesized sound concept, our method proposes to equalize (or homogenize) the different samples. This process consists of extracting the spectral characteristics of a source and target sounds and applying the corresponding spectral transformation.

An analogy of the type of sound transformation presented in this paper is voice conversion, in which a speech signal of a source speaker is transformed to match the characteristics of a target speaker. Voice conversion can be based on probabilistic approaches such as Gaussian Mixture Models, for learning the spectral equalization [2].

2. FILTERING METHODS
2.1 Excitation patterns
One of the most popular models of excitation patterns is the Gammatone filter bank originally proposed by Roy Patterson et al. in 1992 described as Equal Rectangular Bandwidths (ERB, Figure 1). Gammatone filters were conceived as a simple fit to experimental observations of the mammalian cochlea, and have a repeated pole structure leading to an impulse response that is the product of a Gamma envelope and a sinusoid. One reason for the popularity of this approach is the availability of an implementation by Malcolm Slaney, described and implemented as a Matlab Toolbox [3]. The auditory filter bandwidths increase with center frequency, and constitute what is known as the upward spread of masking. The upward spread of masking refers to the fact that low frequency sounds are more effective at masking higher frequency sounds [4].

The auditory filtering concept is in part analog to the constant-Q transform (Brown and Puckette, 1992), but in that case, the bank of filters is of a constant ratio between the center frequency and bandwidth. With the appropriate configuration, the center frequency of the filters of the constant-Q transform can directly correspond to musical notes.

Another approach is based on the works of Stevens, Volkman and Newmann in 1937 that defined MEL, a perceptual unit based on pitch comparisons. We can compose a bank of triangular-shaped filters that symmetrically overlap (Figure 2). We assessed that this approach fits best the use case of environmental sounds that have broadband spectral con-
2.2 Audio transformations based on auditory filtering

Initially, our efforts are focused on reaching a certain homogenization across samples computing the excitation of a MEL filter bank and then processing all the samples to reach a certain and meaningful transformation point.

2.3 Homogenization algorithm

First we generate the filters needed to implement the analysis and filtering stages. After some initial tests, we used 40 MEL-scaled filters to reach a certain compromise between filter overlap and computation time (Figure 2). Then we obtain the excitation of a filter \( m \) in the frequency domain computing the RMS of a windowed frame (we used \( N = 512 \) samples and an overlap of 25% with a hanning window). We make use of the RMS value as the analysis feature that represents the gain contribution of a certain band:

\[
G_m = \left( \frac{\sum_{i=1}^{N} x_i^2}{N} \right)^{\frac{1}{2}}
\]  

(1)

Afterwards we compute the mean for each of the previous excitations across \( n \) frames using \( k \) filters (in the test case, \( k = 40 \)). A normalization of the mean values is carried out in order to compute a weighting that helps to sort the frames by relevance depending on their energy contribution:

\[
W_n = \frac{\text{mean}(Fr_n)}{\text{max}(Fr)}
\]  

(2)

Thus, the weighted gains across frames are computed as:

\[
WG_{m,n} = G_{m,n} \cdot W_n
\]  

(3)

And the source gain excitation for a certain audio file is:

\[
G_{s,m} = \frac{\sum_{m=1}^{k} WG_m}{\sum W_m}
\]  

(4)

Then we compute the mean of the ERB excitations related to a filter \( m \) for all the files in the data set. This is the target excitation we want to reach, and it can be defined as the “homogenization point”. At the end, a scaling factor for each file can be composed as a scaling matrix:

\[
S_{m,\text{file}} = \frac{G_{t,m}}{G_{s,m,\text{file}}}
\]  

(5)

Optionally, a matrix correction can be applied to the scalings, as explained in the next section. Finally the Gain transformations yield from the multiplication of the scaling to the corresponding gain of each filter, that is, the final gain for each filter applied to the homogenized file:

\[
G_{t,m,\text{file}} = G_{\text{source}} \cdot G_{m,\text{file}}
\]  

(6)

2.4 Iterative filtering

The overlap of each of the filters in the MEL filter bank causes a deviation from the computing of the gains. In order to fix this, a gain matrix correction can be applied using a variation of the methods proposed in [4]. Given a gain matrix \( A \) that characterizes the overlap across filters, we can correct the gains as:
$C_g = A^{-1} \cdot \text{Gains}$  

(7)

This method has the issue of computing a correct gain matrix $A$, which is highly tied to the filter type. Moreover, in the case of the gammatone filter bank, there is a noticeable complex influence across bands. In order to deal with the filter overlap we designed another approach based on iterative filtering. We observed that filtering iteratively and computing again the scalings (Figure 4), there is a convergence of the scaling values. We defined a threshold of 2% as exit condition. Depending on the sound, it will take more or less iterations to reach the maximum transformation point (ideally, the spectral envelopes of source and target overlap).

3. TRANSFORMATION EXPERIMENTS

3.1 Samples homogenization

The first use case we considered is the equalization of footstep sounds. We selected a data set of 7 footstep mono recordings, recorded at different levels and containing a sequence of steps on ice, stones and snow from a commercial library of sound effects (East West Blue Box eastwestsamples.com). They are also steps recorded without highly noticeable environmental or background ambience. The overall timbre and spectral envelope is different across samples. We observed that certain homogenization is reached if we compute the overall RMS of the source raw samples and compare it with the resulting processed samples (Figure 5). Moreover, the homogenization is influenced by the differences in levels of the recordings. We also observed that carrying a preliminary segmentation using onset information or manual editing, the homogenization across the different units (in this case, the isolated steps), improves.

3.2 Corpus based texture synthesis

Sound texture synthesis intends to model statistically stationary sounds in order to be able to synthesize target sounds.
which perceptually match analyzed source sounds [6]. Recently, sound texture synthesis has been cast in the framework of data-driven, corpus-based synthesis [7]. In this scheme, units of fixed or variable duration are extracted from the source material and transformed to a multi-dimensional space according to the extraction of perceptual audio features. A feature frequently used as a timbre descriptor are Mel-Frequency-Cepstral-Coefficients (MFCCs). A common problem in corpus-based synthesis method is that often units form clusters in feature space, where the space between clusters is only sparsely populated and the synthesis of smooth transitions between clusters is problematic. One possible way of solving this problem is to artificially expand the feature space by adding variations of existing units to fill the gaps [8]. We applied our transformation method to different recordings of large water streams, also from the Blue Box sample collection. The transformation target is specified by the mean of the energy in each band of the filter bank across all recordings. The corpus is then formed by segmenting each original and transformed recording into texture windows of 600 ms duration and extracting the first 20 mean MFCC coefficients for each resulting unit. Figure 6 shows the projection of the whole corpus of original and transformed sounds onto the first two principal components. The plot shows that the transformed units are located in a region of lower variance along both dimensions, thus making the feature space more dense and more appropriate for synthesizing the centroid of all recordings.

The sample homogenization between source and target was carried out with the following parameters: 512 samples of FFT window length for the filterbank computation and a threshold of 0.02 for the rate of change in the filter gains between iterations, resulting in 6 to 12 iterations, depending on the source and target sounds. Each sound was recorded to one channel, with a sample rate of 44.1 kHz and a resolution of 16 bits.

For this preliminary test we observed and measured slightly better similarity in material and a significant improvement in sensation of recording location (a factor of 0.75 for transformed and 0.39 for non-transformed sounds).

5. CONCLUSIONS

5.1 Analysis of the results
Auditory filtering is a powerful tool for sound transformations of samples, that can also be used in other contexts related to games development like adaptive mixing. Also the implementation of procedural models [10] [12] [13] [14] can benefit from these techniques in order to simplify the sound production process within an audio engine. One of the possible applications of the samples homogenization algorithm beyond the use cases presented, can be the normalization of studio recordings that differ in the microphones or processing gear used at different takes, as well as full music productions that need to compensate the equalization used.

5.2 Future work
The paths for the transformations in the timbre space can be approximated and computed in different ways. A simple linear interpolation has been used for our initial tests, but more sophisticated curve-fitting techniques can also be applied in order to translate the MEL excitation coefficients to different points in the timbre space depending on their content, for instance in low or high frequencies. Also probabilistic approaches such as the Gaussian Mixtures Models mentioned at the introduction could help out to improve the methods developed.

6. ACKNOWLEDGMENTS
We are grateful to the participants that gave us feedback in the listening tests. Thanks to the Freesound project contributions for some of the sounds used in the data set and the initial experiments. The first author holds a grant from the the Spanish Ministry of Education for Master students, the second author is supported by a FI scholarship from the Generalitat de Catalunya.
7. REFERENCES


