Granular Resynthesis for Sound Unmixing

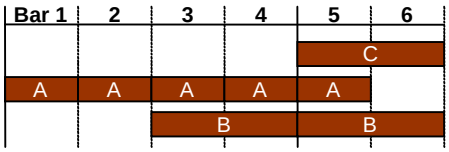
Technical University of Berlin

Gunnar Eisenberg and Thomas Sikora
{eisenberg | sikora}@nue.tu-berlin.de

Communication Systems Group

1. Abstract

In modern music genres like Pop, Rap, Hip-Hop or Techno many songs are built in a way that a pool of small loops, are used as building blocks. These loops are usually one, two or four bars long and build the accompaniment for the lead melody or the singing voice. Very often the accompanying loops can be heard solo in a song at least once. This can be used as a-priori knowledge for removing these loops from the mixture. Based on this knowledge the presented algorithm uses granular resynthesis and spectral subtraction for unmixing the loops.



2. Algorithm

Granular Analysis

Let $x(n)$ be the time signal of the mix-loop and $s(n)$ be the time signal of the subtrahend-loop. The grains are extracted with their anchors having a distance of the hopsize N. They are windowed out of the signals using a Hanning window $w(n)$ of length W.

The grains themselves are denoted with $x_k(n)$ and $s_k(n)$ with k being the grain index. After extracting the grains, they are analyzed using an FFT of size L, yielding in the spectral blocks $X_k(l)$ and $S_k(l)$.

$$x_k(n) = x(n+kN)\cdot w(n) \qquad x_k(n) \xrightarrow{FFT} X_k(l)$$

$$s_k(n) = s(n+kN)\cdot w(n) \qquad s_k(n) \xrightarrow{FFT} S_k(l)$$

For further processing the spectral blocks $X_k(l)$ and $S_k(l)$ are split into magnitude blocks and phase blocks.

- sample rate: 44.1 kHz
- hopsize N = 1024 taps (23,22 ms),
- window length W = 4096 taps (92,88 ms),
- FFT size L = 4096 taps (92,88 ms).

Basic Grain Synthesis

The information gained during the analysis process is now used for setting up the resynthesis grains. The following equations compute the magnitude and phase of each residual grain:

$$\hat{R}_k(l) = \hat{X}_k(l) - \hat{S}_k(l) \qquad \overset{\varphi}{R}_k(l) = \overset{\varphi}{X}_k(l)$$

By these two steps each residual grain's spectral information is now synthesized to:

$$R_k(l) = \hat{R}_k(l)\exp\left(j\,\overset{\varphi}{R}_k(l)\right)$$

From this residual grain spectrum the grain itself can easily be computed by an IFFT. It can then be used for the actual resynthesis in the time domain which is performed by summing up all synthesized residual grains.

The drawback with this basic grain synthesis process is that the spectral character of neighboring grains could vary quite heavily. This results in artifacts which make the frequency with which the grains are placed according to the hopsize, quite audible.

Advanced Grain Synthesis

The advanced grain synthesis technique takes the spectral properties of M neighboring analysis grains with a certain amount into account:

$$\hat{R}_k(l) = \hat{X}_k(l) - \sum_{m=-(M-1)/2}^{(M-1)/2} \hat{S}_{k+m}(l)\,g(m)$$

The factors $g(m)$ which are called shadow factors have a triangular shape. Every grain $R_k(l)$ is transformed back into the time domain by using an IFFT. To avoid block artifacts another less invasive window $v(n)$ is applied to each grain. This window is mainly a rectangular window which fades in and out on the first and last 10% of the window size.

$$R_k(l) \xrightarrow{IFFT} \tilde{r}_k(n)$$

$$r_k(n) = \tilde{r}_k(n)v(n) \qquad r(n) = \sum_k r_k(n-kN)$$

The shadow factors' total window size should be around 185 ms (8192 taps @ 44.1 kHz) to achieve the best resynthesis quality.

3. Results

The algorithm's quality was evaluated with an expert listening test based on a well known mean opinion score (MOS) criterion with 25 musicians as listeners. For each song they were presented the mix signal, the subtrahend signal and the residual signal. The following five songs were used as examples.

No.	Artist	Song
1	Beasty Boys	Hey Ladies
2	Depeche Mode	ASN – Vocals
3	Depeche Mode	Any Second Now (RMX)
4	Electro Nation	Woman Machine
5	Led Zeppelin	Stairway to Heaven

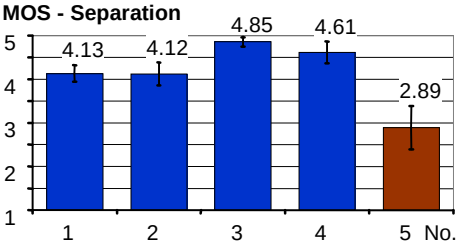
The listeners had to judge two aspects of the sounds. The first aspect was how much from the subtrahend remained in the residual after unmixing, the second aspect was the presence of artifacts in the residual i.e. noise, crackles, fading and musical tones.

For both tests a MOS ranging from one to five was used. The separation's quality was scaled from one meaning "Unsatisfactory (Bad)" to five meaning "Excellent". The artifact impairment was scaled from one meaning "Very Annoying (Objectionable)" to five meaning "Imperceptible".

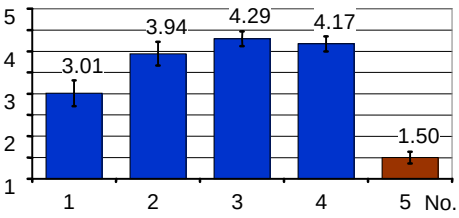
The overall MOS is 4.12 for the separation's quality, which is better than good, and 3.38 for the artifact impairment. This means that artifacts are perceptible and slightly annoying. The separation's quality is better than "Good" for all songs except No. 5. For No. 3 and No. 4 which contain dominant synthetic sounds it is almost "Excellent".

Song No. 5 shows the limitations of the presented algorithm by using two hand played loops. The song's single notes vary heavily in time and the residual signal is a singing voice which reduces the listener's tolerance for artifacts.

MOS - Separation



MOS - Impairment



This research was supported by the European Commission under contract FP6-027026-K-SPACE
Presented at the *9th Int. Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada, September 18-20, 2006