# GRANULAR RESYNTHESIS FOR SOUND UNMIXING

*Gunnar Eisenberg, Thomas Sikora*

Communication Systems Group
Technical University of Berlin
{eisenberg|sikora}@nue.tu-berlin.de

## ABSTRACT

In modern music genres like Pop, Rap, Hip-Hop or Techno many songs are built in a way that a pool of small musical pieces, so called 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 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. In this paper an algorithm based on granular resynthesis and spectral subtraction is presented which makes use of this a-priori knowledge. The algorithm uses two different synthesis strategies and is capable of removing known loops from mixtures even if the loop signal contained in the mixture signal is slightly different from the solo loop signal.

## 1. INTRODUCTION

In the field of musical studio productions and remix applications there exists a high demand for unmixing the single tracks of musical pieces from each other and extracting the used loops, if there are any. An example of a typical track scenario is given in Figure 1. For this scenario extracting the loops would result the loops A, B and C.

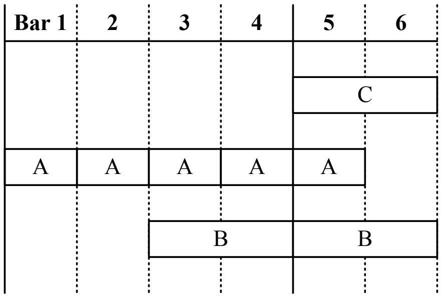

Figure 1: *Typical simplified track scenario for modern music genres.*

Besides the techniques presented for blind source separation so far, this paper presents a straightforward approach for unmixing songs which are built up on a loop base like it is described above. If a loop appears solo in the musical piece its characteristics can be used as apriori knowledge to remove this solo part from the following parts resulting in a residual part which is often a single loop itself. From the scenario shown in figure 1, loop A could be retrieved directly from bar one or bar two. Loop B could be unmixed by removing two bars of loop A from the song's bars three and four. Loop C could be retrieved by removing the newly unmixed loop B and one bar of loop A from the song's bars five and six.

A naiÌĹve implementation for this approach would be to subtract an accompaniment-loop which has been played solo before directly from the mix using the time signals. For this implementation to be successful it would be necessary that the accompaniment-loop is being looped in a sample-identical way. Further the boundaries, especially the starting-sample of the accompaniment-loop would have to be set in a sample-accurate way. Additionally the gain of the loop would have to remain absolutely constant.

These limitations cannot be held for professionally edited studio productions. In professional studio productions subsequently added effects like reverb, equalizers, dynamics, and mastering result in phase and amplitude distortions. Thus the assumption of sample accurate looping does not hold.

Further, many productions do not loop small wave-files directly, which would be classical looping, but repeat only the notes for the loop continuously. Normally these loops are programmed in a MIDI sequencer which sends control information to a synthesizer. The synthesizers itself often produce sounds which are not identical on the sample base. This applies especially to analogue-synthesizers which are being recorded using A/D-converters afterwards. Even digital synthesizers can have this behavior to make the sound livelier.

The mentioned drawbacks which occur for a subtraction using the time signals do not hold for the spectral properties of the tracks to be unmixed. This fact is exploited by the presented algorithm. The method is based on a granular resynthesis process together with spectral subtraction. Although the algorithm does not operate directly in the time domain the nomenclature of this case is used for this paper. This means that the music signal from which the known loop is extracted is called mix, the known loop itself is called subtrahend and the result of this unmixing is called residual. As it is explained in the following in more detail, the spectral properties of the resynthesis signal are adapted using the information which is gained by a granular analysis of the mix signal as well as the subtrahend signal to be removed from the mix signal.

## 2. PREVIOUS WORK

The presented algorithm faces the problem of monophonic sound unmixing with the technique of granular resynthesis. There are several publications on both of the named topics.

Casey has presented a method on independent subspace analysis (ISA) for separating individual audio sources from a single-channel mixture [1]. The method is based on the independent component analysis (ICA) but can operate in scenarios where there are less mixture observations than sources [2].

Virtanen has presented a data-adaptive sound source separation system, which is able to extract meaningful sources from polyphonic real-world music signals [3]. He also has developed an algorithm which transforms each source into a parameterized version which is expressed as a convolution between a time-frequency magnitude spectrogram and an onset vector [4].

Smaragdis has developed a method for the extraction of multiple sound sources from monophonic inputs which is an extension to the non-negative matrix factorization (NNMF). It is capable of identifying components with temporal structure and extracting multiple sound objects from a single channel auditory scene [5].

Since all algorithms developed for mono source separation fail in certain situations the topic has not been solved by now and further research needs to be performed.

The theory of granular synthesis was initially invented by Gabor who proposed that any sound could be decomposed into small acoustical grains [6, 7, 8]. Since then very much research on this topic has been performed under different names. Today aliases of the term “grain" are: “acoustic quantum", “gaboret", “gabor atom", and “wavelet" to name only a few. A good overview is given by Roads who has explained that the potential of granular representations has yet to be fully explored [9]. Further use cases and methods of granular synthesis have been named by Zölzer *et al.* [10].

New approaches to signal analysis by Mallat show several techniques that analytically combine granular synthesis with the broad category of wavelet or atomic decompositions [11].

## 3. ALGORITHM

### 3.1. Preliminaries

Songs to be unmixed using the described algorithm need to fulfill the following requirements:

- The tracks to be removed need to consist of mainly loops.
- The loop to be removed from a mix must appear solo at least once throughout the song.
- The loops of one track must not change their spectral characters heavily from one loop cycle to the next one. This means that for example deep filter sweeps should not occur.

These assumptions easily hold for many songs from modern music genres like Pop, Rap, Hip-Hop or Techno. Furthermore these assumptions directly model almost every modern musical studio production where different tracks and loops are composed in a sequencer or tracker program. Although the presented algorithm only performs the unmixing itself and no automatically loop boundary detection, aligning the song in a sequencer or tracker program will deliver these loop boundaries indirectly.

### 3.2. Granular Analysis

Let $x(n)$ be the time signal of the mix-loop and $s(n)$ be the time signal of the subtrahend-loop. With equation (1) and (2) from both signals grains are extracted with their anchors having a distance of the hopsize $N$. The grains 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 a FFT of size $L$, yielding in the spectral blocks $X_k(l)$ and $S_k(l)$.

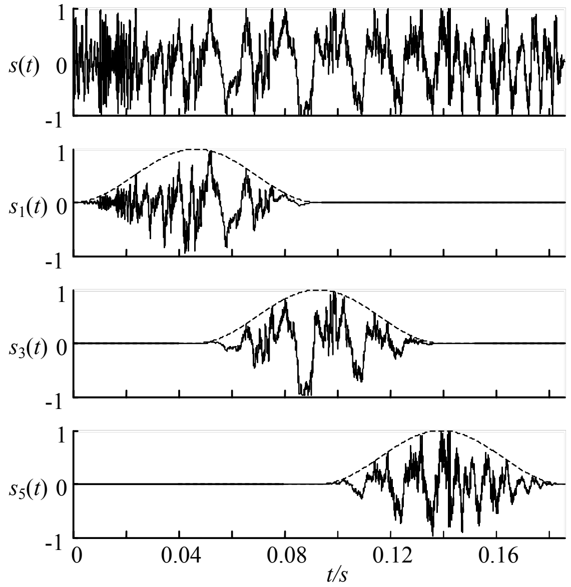

Figure 2: *Extraction process of three analysis grains.*

$$x_k(n) = x(n+kN)\cdot w(n) \quad (1)$$
$$s_k(n) = s(n+kN)\cdot w(n) \quad (2)$$

$$x_k(n) \xrightarrow{\text{FFT}} X_k(l) \quad (3)$$
$$s_k(n) \xrightarrow{\text{FFT}} S_k(l) \quad (4)$$

For further processing the spectral blocks $X_k(l)$ and $S_k(l)$ are split into magnitude blocks $\hat{X}_k(l)$, $\hat{S}_k(l)$ and phase blocks $\overset{\varphi}{X}_k(l)$, $\overset{\varphi}{S}_k(l)$. Figure 2 shows an example-extraction of three grains from the subtrahend signal. For audio signals with a sample rate of 44.1 kHz the following parameters have been determined heuristically to produce good analysis results:

- hopsize $N = 1024$ taps $(23, 22$ ms),
- window length $W = 4096$ taps $(92, 88$ ms),
- FFT size $L = 4096$ taps $(92, 88$ ms).

### 3.3. 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 $\hat{R}_k(l)$ and $\overset{\varphi}{R}_k\ (l)$ of each residual grain:

$$\hat{R}_k(l) = \hat{X}_k(l) - \hat{S}_k(l), \quad (5)$$

$$\overset{\varphi}{R}_k(l) = \overset{\varphi}{X}_k(l). \quad (6)$$

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

$$R_k(l) = \hat{R}_k(l)\cdot\exp\left(j\overset{\varphi}{R}_k(l)\right) \quad (7)$$

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 grain-frequency *i.e.* the frequency with which the grains are placed according to the hopsize, quite audible. Further the subtrahend-loop must match the mix-loop very well. Often this cannot easily be provided since in almost every song there are slight variations from one loop to the next one.

The advanced grain synthesis technique fixes this problem by using more information of neighboring analysis grains to build the spectrum for the resynthesis grains.

### 3.4. Advanced Grain Synthesis

The advanced grain synthesis technique takes the spectral properties of $M$ neighboring analysis grains with a certain amount into account. Therefore equation (5) is replaced by

$$\hat{R}_k(l) = \hat{X}_k(l) - \sum_{m=-(M-1)/2}^{(M-1)/2} \hat{S}_{k+m}(l) \cdot g(m). \qquad (8)$$

The factors $g(m)$ which are called shadow factors have the shape of a window-function which is shown in Figure 3.

For audio signals with a sample rate of $44.1$ kHz the shadow factors' total window size should be around $8192$ taps ($185,76$ ms), to achieve the best resynthesis quality. For a hopsize $N$ of $1024$ taps this means that seven analysis grains are taken into account for equation (8). This results in values for the shadow coefficients which are shown in Figure 3.

The rest of the residual grains' computation does not change, compared to the basic method. This means that equations (6) and (7) are also valid here.

Although the advanced grain synthesis is more robust against spectral changes of the analyzed material it can add an amount of blur to the synthesized grains. This especially occurs in percussive sounds with dominant transients. In these cases the Basic Grain Synthesis is preferred.

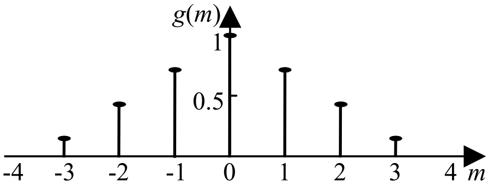

Figure 3: *A typical set of shadow factors.*

### 3.5. Granular Resynthesis

Every grain $R_k(l)$ is transformed from the spectral domain back into the time domain by using an IFFT.

$$R_k(l) \xrightarrow{\text{IFFT}} \tilde{r}_k(n), \qquad (9)$$

Since we have manipulated the amplitudes of the contained sinewaves it is not guaranteed that the grain still has the shape of the Hanning window which we used for the initial extraction process. To avoid block artifacts which may occur when the grains are summed together 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. This window should have the same length as the Hanning window which was used for the analysis grain extraction.

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

Table 1: *Demo songs used for evaluation.*

$$r_k(n) = \tilde{r}_k(n) \cdot v(n). \qquad (10)$$

The actual resynthesis of the output signal $r(n)$ is performed by summing up all edge-corrected grains $r_k(n)$ at their correct time position:

$$r(n) = \sum_k r_k\left(n - kN\right). \qquad (11)$$

## 4. RESULTS

The algorithm performs a psycho-acoustically motivated sound separation based on spectral properties. This makes it almost impossible to evaluate the algorithms quality in a purely deterministic way. Therefore the algorithm's quality was evaluated by performing an expert listening test based on a well known mean opinion score (MOS) criterion [12]. The five songs depicted in figure 4 were used as examples. The songs come from the genres Rap (No. 1), Pop (No. 2, No. 3) and Techno (No. 4). To show the limits of the algorithm also one Ballad was processed (No. 5) which does not hold the preliminaries named in section 3.1. The sound demos together with more examples can be found at our institute's website[1].

The MOS listening test has been performed with 25 musicians. For each song they were presented the mix signal, the subtrahend signal and the residual signal. For the unmixing the parameter set proposed in section 3 was used. These settings focus more on a good separation's quality than on little artifacts. This is important since these two quality aspects somewhat contradict each other.

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 results of the MOS test are depicted in figure 4. The overall MOS is 4.12 for the separation's quality, which is better than

[1] www.nue.tu-berlin.de/wer/eisenberg/unmixing/

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".

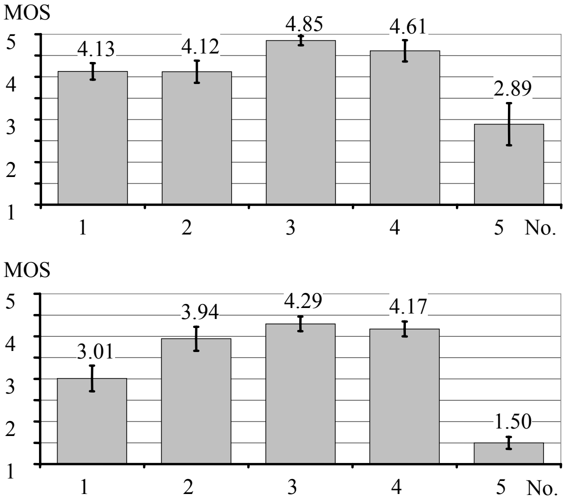

Figure 4: *Mean values of the MOS test for the separation's quality (top) and artifact impairment (bottom).*

The artifact impairment depends on the nature of the processed material. For songs which sound very synthetic like No. 3 and No. 4 the sound artifacts occurring in the residual signal are almost imperceptible. If the residual signal is a singing voice like in No. 1 and No. 2 the artifacts are just perceptible and slightly annoying. If there are little differences between the notes played in the subtrahend and in the mix, like in the bass line of No. 2, these cannot be removed completely and remain present in the residual, although this is hardly audible.

Song No. 5 shows the limitations of the presented algorithm by using a hand played loop as subtrahend which does not hold the limitations mentioned in section 3.1. The song's single notes vary heavily in time and have different spectral characteristics. Furthermore the residual signal is a singing voice which reduces the listener's tolerance for artifacts. After unmixing the residual signal contains annoying artifacts and the unmixing has almost a fair overall quality.

## 5. FUTURE WORK

Besides the expert listening sessions performed for the presented paper, a MOS test with more examples and more listeners is prepared to illustrate the performance of the method.

The songs which the algorithm can unmix contain a high amount of self-similarity by default. The next step for the presented system will be to evaluate this self similarity and use it to perform automatic boundary detection for loops. This would result in less manual work for the user of the system since the system would present reasonable suggestions.

## 6. REFERENCES

[1] M. A. Casey, "Separation of mixed audio sources by independent subspace analysis," in *Proc. Int. Comp. Music Conf. (ICMC'00),* Berlin, Germany, 2000, pp. 154–161.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[3] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. Int. Comp. Music Conf. (ICMC'03),* Singapore, 2003, pp. 231–234.

[4] ——, "Separation of sound sources by convolutive sparse coding," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2004, paper 55.

[5] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Int. Cong. Independent Component Analysis and Blind Signal Separation (ICA)*, 2004, pp. 494–499.

[6] D. Gabor, "Theory of communications," *J. Inst. Elect. Eng.*, vol. 93, no. III, pp. 429–457, 1946.

[7] ——, "Acoustical quanta and the theory of hearing," *Nature*, vol. 159, no. 4044, pp. 591–594, 1947.

[8] ——, "*Lectures on communication theory*, Technical Report 238, Research Laboratory of Electronics, Cambridge, Massachusetts, Massachusetts Institute of Technology," 1952, 238.

[9] C. Roads, *Microsound*. Cambridge, Massachusetts: MIT Press, 2002.

[10] P. Dutilleux, G. D. Poli, and U. Zölzer, "Time-segment processing," in *DAFX – Digital Audio Effects*, 2002, pp. 201–236.

[11] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego: Academic Press, 1998.

[12] N. S. Jayant and P. Noll, *Digital Coding of Waveforms – Principles and Applications to Speech and Video*. New Jersey: Prentice-Hall, 1984.