

COVER PAGE

Document downloaded by @DAEL

Sat Nov 28 00:18:24 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Note-Based Sound Source Separation in Monoaural Polyphonic Music, Kristóf Aczél and István Vajk, Acta Acustica **vol. 96** (Number 5), 2010, pp. 947-958

DOI
https://doi.org/10.3813/AAA.918353

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 96 (2010) 947 – 958
DOI 10.3813/AAA.918353

Note-Based Sound Source Separation in Monoaural Polyphonic Music

Kristóf Aczél, István Vajk
Department of Automation and Applied Informatics, Budapest University of Technology and Economics, 3-9. Műegyetem rkp., 1111, Budapest, Hungary aczelkri@aut.bme.hu, vajk@aut.bme.hu

Summary
Decomposing a polyphonic musical recording to separate instrument tracks or notes has always been a considerable challenge. Such a signal is the superposition of many separate tracks, and it is currently practically impossible to extract the component tracks without the information that was lost at the superposition. This paper introduces a new way of sound separation of mono-aural digital recordings. Complete system architecture is proposed that recovers the lost information by using a model of real instruments in order to make the separation of individual musical notes possible. The separation method targets the processing and correction of musical recordings that cannot be retaken.

PACS no. 43.75.Zz, 43.60.Uv

1. Introduction

This research is motivated by the problem of the correction of existing musical recordings by fixing bad notes or altering the polyphonic structure in other ways like note pitch shifting, volume controlling, etc. A musical recording is typically a polyphonic material consisting of many individual notes originated from a number of instruments. A recording is taken using one or more microphones. Although in an ideal case each instrument has its own, in real life microphones are usually assigned to groups of instruments. Audio signals are then downmixed typically to two channels in the case of a CD-quality target signal.

After this mixing process, no correction can be made to the individual notes in the recording. Any kind of filter can only be applied to the whole recording including all the other notes that are simultaneously played at a specific time. Our long term interest in sound separation is motivated by the problem of correcting recordings that have incorrect notes (either in frequency, volume, intonation, length, etc).

This paper concentrates on the separation of the signal of an arbitrary user-specified note from the recording. The separation of note signals other than the selected one is, although favored, not in the focus. This makes the approach particularly suitable not only for filtering out bad notes, but doing any other kind of transformation to the recording, e.g. changing the melody played by one of the instruments while leaving the others intact.

Reliable automatic musical transcription and instrument recognition is out of reach of today's technology. Thus, in our work we allow a reasonable amount of user input and processing time to achieve better separation quality. User input involves entering the musical score (note onset/offset, frequencies, used instruments). Due to the nature of real-life music, this input will never be 100% accurate, even if the user is presented with some kind of hint about the concrete recording (e.g. spectrogram). However, it can be precise enough for getting a first rough estimate on the note parameters in the recording, instead of relying on musical transcription [1, 2, 3] and music instrument recognition (MIR) [4, 5, 6] algorithms, which are currently inferior to the performance of a human listener.

The work does not focus on noise reduction of either the recording or the separated notes. Any noise contained by the original recording may appear on the output of the separation algorithm. If noise is a concern, standard noise reduction methods [7, 8] should be applied to the recording prior to the separation.

There is a great need for complementary information in sound separation in addition to the raw sound signal that is being processed. The complexity of separation lies in the fact that the information we would like to retrieve from the original signal is actually not present. Significant amount of research work has been devoted to regaining the lost information in different ways.

Model based systems represent a very promising approach. In this category a parametric model of the input sources is established that serves as a set of constraints on the output signals. The model parameters are estimated by observing the mixture signal. The two main branches of this approach are rule-based algorithms [9, 10] that use implicit prior information for building the models, and Bayesian estimation [11, 12], where prior information is explicitly given using probability density functions. In mu-

Received 4 January 2008,
accepted 21 April 2010.

© S. Hirzel Verlag · EAA

sic applications, the most commonly used approach is sinusoidal modeling which suits separation of pitched instruments and voiced speech very well. [13, 14] describe a method that separates harmonic sounds by applying linear models for the overtone series of input sound signals. The method is based on a two-stage approach: after applying a multipitch estimator to find the initial sound parameters more accurate sinusoidal parameters are estimated in an iterative procedure. Separating the spectra of concurrent musical sounds is based on the spectral smoothness principle.

Unsupervised learning methods usually operate on the basis of simple non-parametric models, and require less information on the original sources. They try to gather information on the source signal structures from the mixed data itself using information-theoretical principles, such as statistical independence between the sources. The most common approaches used to estimate the sources are based on independent component analysis (ICA), non-negative matrix factorization (NMF), and sparse coding. These algorithms usually factorize the spectrogram (or other short-time representation of the signal) into elementary components. This is followed by clusterization that builds the output separated channels from the elementary components.

ICA [15, 16, 17, 18] is employed in many blind source separation systems, where no information on the source signals is available. One of the most important scenarios of this technology is multichannel source separation, where an (often large) set of observers are employed to record an event. The travel time of the signal and the difference of the numerous recorded signals are used to increase the receiver sensitivity in the direction of signals of interest and to decrease sensitivity in the direction of other sounds or noise.

NMF [19, 20, 21, 22] methods exploit the fact that the mixture signal is the composition of basis vectors, with non-negative coefficients. Each of these vectors represent a part of a single sound source. In the signal model the non-negativity restriction alone is a sufficient condition for the separation of the sources in most cases.

Sparse coding [23, 24] represents the signal as a linear combination of a set of source signals where only a few sources are active at a time, thus the probability that a candidate signal has zero gain is high. The approach is efficient for learning structures and separating sources from mixed data.

Another interesting approach is introduced in [25]. The method performs the task of separation based purely on azimuth discrimination within the stereo field.

This paper proposes a separation system that is based on the model-driven approach. A global architecture is given for the system, while its parts are also discussed in detail. The established model of the sources is introduced (instrument prints), and the Simplified Energy Split (SES) algorithm is proposed for the task of redistributing the energy of the original mixture. The novelty of the approach lies in both the way instruments are modeled, and the SES algorithm, which makes it capable of separating sources even when they share the same fundamental frequency.

The following sections guide the reader through the separation process, each section covering one important block of the algorithm. Section 2 gives an overview of the blocks and the signal flow in the system. Section 3 describes the instrument model that is used by the separation process. Section 4 presents the description of the source separation problem. The separation needs the properties of the musical notes in the recording. While some of the properties can be specified by the user, some have to be estimated algorithmically. This is covered in section 5. Section 6 introduces the most important block in the system, the Simplified Energy Splitter, which carries out the actual source separation, while section 7 covers a method for handling the beating artifact that may appear in the separated output signals. Section 8 deals with synthetic tests and measurement results. Comparison to other approaches is covered in section 9. Finally, section 10 concludes.

2. Overview of the separation process

This section gives an overview of the separation process. The building blocks are shortly described, while the details of each block are elaborated in later sections.

The whole separation system operates in frequency domain. For that reason all time-domain signals have to be transformed to frequency domain before processing. Likewise, at the output of the system the frequency-domain images of the separated note signals must be transformed back to time domain. In addition to the usual STFT transformation the former conversion also includes a frequency estimation step that is based on the method proposed by Brown [26]. It makes it possible to obtain a far more precise spectrogram than the one provided by the STFT algorithm by calculating the instantaneous frequencies of the STFT bins.

The system can operate in two modes. The first one is the instrument print creation mode. Here the system takes sample waveforms from real-life instruments and transforms them to a representation that will later be useful for separation purposes. For this purpose we propose the instrument print model that is based on the *bandogram* representation of instrument sounds [27]. A bandogram is similar to a spectrogram in many respects. The main difference lies in its resolution through the frequency range. A bandogram stores the sums of energies located in frequency bands of logarithmically growing width. For separation purposes we will need sets of instrument bandograms from each instrument playing in the musical piece to be separated. Section 3 elaborates the details of bandogram and instrument print creation. Figure 1 depicts the signal flow and blocks of the process, while the notation used is explained in Table I.

The second operation mode of the system, the separation mode, is depicted in Figure 2. It allows for the extraction of individual note signals from the source recording into multiple output tracks. In this operation mode the system has three inputs: the original music, the musical score

Table I. Notation of the sound separation block diagram.

W	*Waveform:* Simple waveform
S	*Spectrogram:* Simple FFT spectrogram. Stores the $c_{k,t}$ amplitude and $\varphi_{k,t}$ phase values for each bin.
F	*FE spectrogram:* Frequency estimated spectrogram. $c_{k,t}$ bin amplitudes and $\varphi_{k,t}$ phase values remain the same as those of the simple FFT, but an $f_{k,t}^{inst}$ instantaneous freq. value is stored in addition for each bin.
B	*Bandogram:* A spectrogram split to subbands, in which the energy is summed. Only these sums are stored, no detailed information on bin amplitudes and no phase information either.

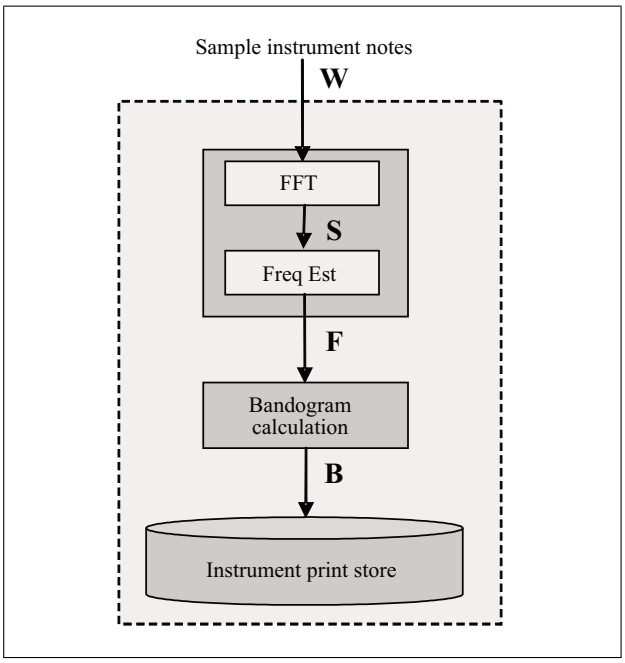

Figure 1. Signal flow and block diagram of the instrument print creation process.

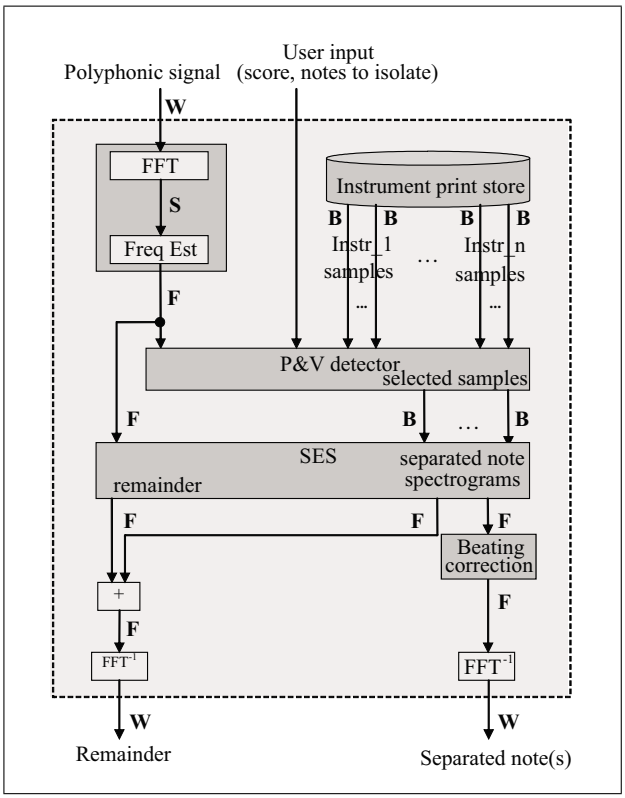

Figure 2. Signal flow and block diagram of the separation process.

that is entered by the user, and the instrument prints that were created in the first operation mode. The difficulty of the separation task can be attributed to the lack of information about the original notes of the recording. Section 4 covers the details of the separation problem.

The system has three main parts in the second operation mode. These are: the playmode & volume (P&V) detector, the Simplified Energy Splitter (SES) and the beating corrector. The task of the P&V detector is to complete the information on the musical score that was given by the user. It was found, that the user input concerning the locations of the notes can be considered satisfactory both in time and frequency. However, there are several ways to play the various instruments. Hitting the keys of the piano with different strength produces different variations of the same note, while one produces this effect with the saxophone by blowing the instrument in different ways. To tackle this problem we introduce the notion of the playmode that addresses this property of instruments collectively. In early tests it turned out that specifying the correct playmode is often difficult for the user. As later steps require this information, the P&V detector is employed to gather the missing data. It selects the most suitable bandograms from an instrument print for each note in the recording at a given time. This is covered in section 5.

The Simplified Energy Splitter can be considered the main module in the separation architecture. Its task is to redistribute the energy content of the input recording between the several output channels. It provides a fair division of the energy between the channels based on the notes' onset and offset, frequency function, volume and playmode function. The detailed description of this block can be found in section 6.

The output signals generated by the SES may contain amplitude fluctuations on the fundamental or overtone frequencies of the separated notes. In other words, the output of the SES may suffer from the beating effect. This is due to the fact that the amplitude fluctuations were already present in the source recording. However, while they may have been unnoticed by the listener due to the complexity of the sound mixture, they become audible in the separated note signals. This phenomenon is addressed in the last step

of the separation process, the beating correction, which is covered in section 7.

The following chapters will guide the reader step by step through the separation process. Each section will cover one block in detail. Due to size limitations, however, some of the blocks will not be exhaustively discussed in this paper.

3. Instrument prints

The main complexity of sound separation lies in the paradox that we need to regain information from a signal that does not fully contain it. Even today we are far from understanding every detail of human sound perception and selective listening. However, studies conclude that we use the memory of instruments and the musical score of the piece being performed [28, 29]. This is our source of additional information.

At some point we will have to supplement additional data to the separation system in order to complete the missing data. Copying nature has many times been proven to be the right approach. This section presents a way of implementing a memory of known instruments that tries to mimic the way the human brain works.

Our approach for the separation is halfway between model based and sparse approaches. On the one hand we have established a vector model with adjustable parameters to describe different instrument note signals. This model stores an effective representation of individual instrument note signals. On the other hand, the stored vectors can also be considered a dictionary built of different samples from different instruments, from which only a few are active at any given time instance. From this perspective the approach is also a sparse representation of polyphonic music.

Various dictionaries have been used in the literature. Windowed sinusoidal atoms were used for building dictionaries in speech modification [30] and audio coding tasks [31]. Gabor atoms, the complex expansion of sinusoidal atoms have been used in audio signal decomposition scenarios [32] and audio coding [33]. Damped sinusoids were used in [34], and wavelet-based approaches were discussed in [35]. Chirped Gabor atoms were proposed in [36] and [37], on the ground of which [38] built a mid-level representation of instrument sounds. Latter defined instrument specific harmonic atoms and molecules for analyzing musical instruments that produce harmonic sinusoidal partials only.

In [39] experiments are presented that examine the dynamic attributes of timbre evaluating the role of onsets in similarity judgments. In these experiments human listeners were asked to compare samples of real instruments that were artificially modified in various ways. It also gives an overview of other researches tackling the challenges of the identification of the most important instrument sounds properties that allow a human listener to di stinguish them. The presented experiments lead to the conclu sion that the rapidity of the attack of instrument tones as well as general differences in their amplitude envelopes contribute the most to similarity judgments. Our instrument model is partly based on the results of [39] in the sense that it contains feature vectors based upon the amplitude envelope of instruments. The model proposed here also shares common features with the mid-level representation described in [38], such as the way the amplitude envelope is calculated (by summing the w indowed sinusoidal partials).

In our research, however, we face a slightly different challenge. Since we aim to resynthesize, not only recognize the individual note signals, our separation scenario requires a more detailed model of the instruments than recognition tasks. We also consider instruments with non-harmonic components, which requires knowledge also on the aperiodic part of instrument signals. On the other hand, in our case many parameters (onset, offset, fundamental frequency, instrument type) of the individual notes in the recording are available prior to the separation. Therefore we need a slightly different model than the one used in [38].

Our instrument model, the *instrument print*, contains samples from an instrument of different frequencies and with different *playmodes*. The term *playmode* refers to the way the instrument was played, e.g. the hardness of a piano key hit, the blowing strength of the flute or the intonation of a saxophone note. One print can have more than one playmode dimension, depending on the ways an instrument can be played. There can be e.g. a *hardness* dimension for piano notes ranging from 1 to 10, while a saxophone may have *loudness*, *sharpness* etc. dimensions. These dimensions cannot always be defined in mathematical terms; very often they can only be labeled by subjective ones like *loudness*, *sharpness*, *warmth* etc. In short, an instrument print is much like a collection of samples of different f_0 fundamental frequencies and different values in the playmode space $\boldsymbol{A} = [m_1, m_2, \ldots, m_p]$. Although it is not yet the final definition, for now we can regard a print as an $\boldsymbol{A}$ amplitude function

$$\boldsymbol{A}(\boldsymbol{M}, f, f_0, t), \tag{1}$$

with the following conditions:

$$t, m_x, f_0 \in \mathbf{R}^+, \qquad 0 < m_x < m_{x,\max},$$
$$0 \le t < \infty, \qquad 0 < f, \quad f_0 \le 20000\,\mathrm{Hz}.$$

This function shows how amplitudes change over time t over the frequency range f for a specific note played on f_0 fundamental frequency and played with playmode $\boldsymbol{M}$.

In reality, we will not have all the samples an instrument can produce, only a few of them, located on different frequencies and playmodes. Our samples will also be finite in time. As an instrument may produce slightly different variations of the same note each time it is played, it is also advisable to tolerate small differences of the note signals. Therefore, instead of storing a full spectrogram, we divide the frequency range to subbands and only store the sum of the amplitudes in these subbands. This tolerance of the model will be important later in the process

of matching the spectral envelope of the stored samples to the ones in the real recording. Too wide subbands lead to a model where the distinction of the various instruments is not satisfactory, while too narrow bands render the system too sensitive to small spectral differences of the note signals.

Our representation will be referred to as a *bandogram*. Each bandogram is calculated from a sound signal that contains exclusively one note originating from the specific instrument. The subbands are aligned on a logarithmical frequency scale. A bandogram can be defined as

$$A_{\boldsymbol{M},f_0,b,t} = \sum_{k\in\rho(t,f_0,b)} c_{t,k}, \tag{2}$$

where $c_{k,t}$ is the amplitude value of bin k at time t, b identifies a specific subband where the amplitudes are summed, R defines the resolution in frequency, that is, the number of subbands per octave, and $\rho(t, f_0, b)$ identifies a set of bins whose calculated instantaneous frequency is in b subbands distance from f_0:

$$\begin{aligned} &K \in \rho(t, f_0, b) \quad \Leftrightarrow \\ &f_0 \cdot 2^{(b-0.5)/R} < f_{t,k}^{\text{inst}} < f_0 \cdot 2^{(b+0.5)/R}. \end{aligned} \tag{3}$$

In our separation system we used $R = 12$ as it provides good enough resolution in frequency on a logarithmical scale, and it is also easy to understand as an octave consists of 12 semitones. With this choice the value of b becomes the distance of a bin's frequency from f_0 in semitones.

The number of recorded instrument bandograms is finite both in frequency and playmode spaces. However, the separation step (that is described in detail later) will require samples of virtually any frequency and playmode. Missing samples are linearly interpolated from the existing ones.

4. The separation problem

In this section we have a short discussion about the complexity of the source separation problem. We will find that due to the lack of information alternative solutions have to be considered that approximate the solution of the original equations. Thus, a simplified version of the equation system will be proposed further on which makes it possible to carry out the separation at the expense of slightly lower quality. By applying the change the separation problem will be simplified to a credit assignment issue.

Let $\boldsymbol{c}_{r\tau} = \{c_{r\tau,k}\mathrm{e}^{\mathrm{j}\gamma_{r\tau,k}}\}$ denote the spectral image of the polyphonic recording at time $r\tau$ ($r \in \mathbf{N}$), $\boldsymbol{s}_{i,r\tau}^{\text{orig}} = \{s_{i,r\tau,k}^{\text{orig}}\mathrm{e}^{\mathrm{j}\sigma_{i,r\tau,k}}\}$ represent the spectral image of original note i, and $\boldsymbol{d}_{r\tau} = \{d_{r\tau,k}\mathrm{e}^{\mathrm{j}\delta_{r\tau,k}}\}$ denote the additional noise component. The separation problem can be formed as

$$\boldsymbol{c}_{r\tau} = \sum_{\forall i} \boldsymbol{s}_{i,r\tau}^{\text{orig}} + \boldsymbol{d}_{r\tau}, \tag{4}$$

where $c_{r\tau,k}$, $s_{i,r\tau,k}^{\text{orig}}$, $\sigma_{i,r\tau,k}$, $\gamma_{r\tau,k} \in \mathbf{R}^+$. Equation (4) is an underdetermined system of equations which cannot be solved unambiguously without any further constraints. Unfortunately, our knowledge on the original notes is rather limited. No precise information is available on the fundamental frequency $f_{0,i}(r\tau)$ of the original notes, their onset/offset, and their playmode is even more uncertain, not to mention that it may also change over time. Each original note will 'resemble' one instrument sample in our database to some extent, but there are no perfect match es ever. This lack of constraints does not allow for decomposing the recording $\boldsymbol{c}_{r\tau}$ to an array of signals that are perfect replicas of the original $\boldsymbol{s}_{i,r\tau}^{\text{orig}}$ ones. As the exact solution to (4) is very complex, we have to set a more reasonable target: decomposing the recording to signals which resemble the original components or – lacking the original note signals – resemble at least samples of the same instrument.

The term *resemblance* is of course a natural expression taken from real life rather than an exact mathematical expression or measure that can be calculated. In the case of an automated algorithm, however, it must be defined in an exact manner in order to be able to interpret and validate the results of the separation step. There is a difference between cases where the original note signals are available, and can be compared to the separated note signals (e.g. test scenarios), a nd cases where the separat ion results can only be compared to instrument prints. Section 5 provides more details on the latter, defining the *resemblance* of separated notes to bandograms, while section 8 covers test cases, where the output quality of the separation system can be measured and compared to other separation approaches.

As (4) cannot be solved, we have to reach for alternative solutions or simplifications to the original problem. It is advisable to modify (4) in a way that causes little difference in the perceived quality by humans, while allowing for a simpler solution for (4). Many efforts have been dedicated to studies of the human auditory system. It was found that human auditory perception is insensitive to the phase information of the component sign als, provided the components are not closer to each other than a *critical bandwidth* [40]. Many researches conclude that hearing is based on the amplitude of the frequencies, and is *generally* insensitive to their phase [41]. They state that the shape of the time domain waveform is only indirectly related to hearing, and usually not considered in audio systems. Subjective evaluations have shown that the relevancy of phase parameters of sinusoids generally is so low that a representation that neglects the phase information can very effi ciently provide the same perceived quality as the original signal if phase continuity is guaranteed between subsequent frames [42].

Based on these findings, we propose a simplification for that eliminates the unknown $\sigma_{i,r\tau,k}$ and $\delta_{r\tau,k}$ phases from the original equation system. Let's assume that

$$\gamma_{r\tau,k} = \sigma_{i,r\tau,k} = \delta_{r\tau,k} \tag{5}$$

for all instruments (i), time frames ($r\tau$) and bins (k). This modification, motivated by the characteristics of human audio perception, exploits the fact that the human ear is

much more sensitive to amplitude differences than phase distortions. The simplified equation system can be formed now as

$$\left| c_{r\tau,k} \right| = \sum_{\forall i} \left| \hat{s}_{i,r\tau,k} \right| + \left| \hat{d}_{r\tau,k} \right|, \tag{6}$$

where $\hat{\boldsymbol{s}}_{i,r\tau} = \{\hat{s}_{i,r\tau,k} e^{j\gamma_{r\tau,k}}\}$ and $\hat{\boldsymbol{d}}_{i,r\tau,k} = \{\hat{d}_{r\tau,k} e^{j\gamma_{r\tau,k}}\}$ represent the amplitude spectrum of the separated instrument note signals and the noise (residual of the separation). Now we are looking for $|\hat{s}_{i,r\tau,k}|$ and $|\hat{d}_{r\tau,k}|$ for each i, $r\tau$, k if $|c_{r\tau,k}|$ and $\gamma_{r\tau,k}$ are known.

Of course, any modifications will have greater or smaller impacts on the quality of the separation, causing artifacts in the output signals which may have to be addressed later. The simplification proposed here has the following drawbacks:

- It does not handle periodic signals with closely located frequencies well. Two or more such signals cause beating in the original recording: the signal will seem to contain energy only on one frequency with fluctuating amplitude. If this effect is not eliminated, the same cancellations will be present in the separated note signals too.
- A periodic signal and an aperiodic one in the same frequency range will not be separable. Either there will be both periodic and aperiodic components in the separated notes, or one of them will be zero, while the other contains both components.

Experiments and mathematical calculations have been carried out also to determine the energy loss ratio caused by the simplification. The average values showed close to 0% of energy loss for distant frequencies, and 35% for very close frequencies. In other words, in a typical case 0% and 35% will be lost, respectively, of the original energy if the simplified algorithm is employed. However, as we will see later, if this energy loss can be detected, it is possible to restore some of the lost energy to the separated notes.

5. Playmode and volume detection

This section covers the first step of the note signal separation process. While the user can specify the location of the instrument notes in frequency and time, he may not be capable of entering either the playmode matrix $\boldsymbol{M}_{i,t}$ or the volume. The automatic detection of the playmode and volume will be covered further on. The detection proposed here uses the actual separation step that is described in detail in the next section.

The task of the P&V detector is to find an optimal playmode matrix $\boldsymbol{M}_{i,t}$ that results in the least possible separation error later in the SES source separation step. Since we had to give up on the perfect reconstruction of the original notes in the recording, separated notes will always deviate from their original. This deviation can be measured in test scenarios where we have the original notes as a reference, but in a typical case it is not possible to calculate the exact separation error, thereby leaving us no other alternative than comparing the separated note signals to the existing samples instead of the original note signals.

The nature of this comparison requires a measure, a *resemblance score* that corresponds to the operation of the human ear and cochlea. It should allow deviances that are not audible to the human ear, and punish in the cases where a human listener would notice the difference. Based on the findings of section 3 we propose the following definition of the separation error:

$$\mathrm{err} = \beta \mathrm{Err}_{\mathrm{note}} + (1 - \beta)\mathrm{Err}_{\mathrm{rem}}, \tag{7}$$

where $\mathrm{Err}_{\mathrm{note}}$ stands for errors in the separated note signals, showing how much they resemble the prints used in the separation, and $\mathrm{Err}_{\mathrm{rem}}$ represents the errors in the remaining coefficients, that is, the remaining amplitudes in the amplitude spectrum of the original signal that could not be associated with any of the output note signals. β represents the preference between the two. Exploiting the fact that a human listener will typically notice amplitude and frequency deviation easily while being less sensitive to phase deviations the two error measures can be defined in the following way:

$$\begin{aligned} \mathrm{Err}_{\mathrm{rem}} &= \sum_{r=0}^{\infty} \sum_{\forall k} \hat{d}_{r\tau,k}, \\ \mathrm{Err}_{\mathrm{note}} &= \sum_{r=0}^{\infty} \sum_{\forall i} \hat{u}_{i,r\tau}, \end{aligned} \tag{8}$$

with

$$\hat{u}_{i,r\tau} = \sum_{\forall b} \left(A_{i,\boldsymbol{M}_i,f_0,b,r\tau} - \sum_{k \in \rho(r\tau,f_0,b)} \hat{s}_{i,r\tau,k} \right), \tag{9}$$

where $\hat{u}_{i,r\tau}$ represents how the bandogram in the instrument print differs from the sum of amplitudes of the separated note for note i at time $r\tau$. The definition focuses only on the sum energy and its distribution over the frequency range. It allows the spectrogram of the separated note to deviate from the sample to some extent on the magnitude axis and also on the frequency axis (e.g. when the separated note has a slightly different fundamental frequency (f'_A) from its parent sample f_A). Although it contains no explicit constraint on the phases, in our case phase continuity between subsequent time frames is guaranteed due to the nature of the separation algorithm, as it preserves the original phase information of the input signal.

The above defined separation error is to be used in the calculation of the optimal playmode matrix. For the sake of convenience the volume is also incorporated in $\boldsymbol{M}_i$ from now on. $\boldsymbol{M}_i$ is by definition perfect if the separated note signals are the perfect replicas of the parent instrument samples that were used in the energy split ($\mathrm{Err}_{\mathrm{note}} = 0$), and the residual of the separation is zero ($\mathrm{Err}_{\mathrm{rem}} = 0$). In general, matrix $\boldsymbol{M}_i$ is considered good if an energy split step that uses matrix $\boldsymbol{M}_i$ generates note signals that resemble their parent sample while letting the remainder get as close to zero as possible.

The most obvious way of finding the optimal $\boldsymbol{M}_i$ is to iterate through all the possible combinations. The optimum can be found by carrying out a separation step for each combination, always measuring the separation error. Depending on the size and possible values of $\boldsymbol{M}_i$ the number of combinations needed for finding the best combination of the instrument bandograms may require huge computational power. If we consider that the playmode space is continuous in reality, it is not even possible to iterate through all the combinations. However in practice it was found that close-to-optimum results can be achieved by trying only 5 levels in each playmode dimension. In this case the search requires $O(5^I)$ steps, where I is the number of concurrent instrument notes. Taking into consideration that our objectives do not include real-time operation, and in a typical polyphonic piece I is limited, the simple brute-force approach fits our needs in most cases.

6. The simplified energy splitter

This section describes the heart of the separation process, the Simplified Energy Splitter. The SES has the task of redistributing the energy in the source recording between the output channels, the separated note signals.

In this step the bandograms selected by the P&V detector will be used to recreate the spectrograms of the note signals that are being separated from the remaining part of the recording. Assuming that we know the exact frequency, volume and playmode of a certain note that we want to separate, we propose the following iterative algorithm to divide the energy between the target note signals. We start out with the FE amplitude spectrum of the original recording. We complete the separation in D steps. In each step a fraction of the amplitude in the selected bandograms is transferred from the amplitude spectrum of the recording to the separated note signals. In an ideal case we will be able to transfer the required amplitude fraction in each step. However in a usual case the remaining amplitude in the original signal will decrease to 0 before we could reach the last step. This is due to the cancellations in the original recording (see section 4). Nevertheless, this strategy ensures fair division of the amplitudes in the recording between the output note signals. Having D steps avoids cases where the whole amplitude would get transferred to one output channel, leaving no amplitude for the other.

Let $\hat{\boldsymbol{c}}_{[0,0],r\tau} = \{\hat{c}_{[0,0],r\tau,k}\} = |\boldsymbol{c}_{r\tau}|$ denote the initial amplitude spectrum of the recording. Let $\hat{\boldsymbol{s}}_{i,[d],r\tau}$ denote the current amplitude spectrum of the separated notes, with $\hat{\boldsymbol{s}}_{i,[0],r\tau} = 0$ initially. Each step $[d]$ $(0 \le d < D)$ comprises I substeps $(1 \le i \le I)$, where I denotes the number of instruments in the time frame. Each substep transfers an $\alpha_{[d,i],b,r\tau}$ fraction of the sum amplitudes of the reference bandogram A_i (in each of its b subbands) from the current remaining $\hat{\boldsymbol{c}}_{[d],r\tau}$ to the separated ith note $(\hat{\boldsymbol{s}}_{i,[d],r\tau})$ if still possible. $\alpha_{[d,i],b,r\tau}$ can be calculated as

$$\alpha_{[d,i],b,r\tau} = \frac{A_{i,M_i,f_0,b,}(r\tau - T_{\text{onset},i})}{D}, \tag{10}$$

where $T_{\text{onset},i}$ is the onset of note i. Inside one substep $\hat{\boldsymbol{c}}_{[d,i],r\tau}$ denotes the amplitudes left in the recording during the execution of step $[d]$ after transferring the amplitude fractions to notes $1 \ldots i-1$. In each substep we try to subtract $\alpha_{[d,i],b,r\tau}$ from this remaining amplitude spectrum and add it to the output signals. If the recording still has enough amplitude in the bins in question, that is

$$\sum_{k \in \rho(r\tau, f_0, b)} \hat{c}_{[d,i],r\tau,k} \ge \alpha_{[d,i],b,r\tau}, \tag{11}$$

then $\alpha_{[d,i],b,r\tau}$ can be subtracted from the bins of the recording, and added to the bins of the output signal. In this case the next $c_{[d,i+1],r\tau,k}$ values are calculated in the following way:

$$\hat{c}_{[d,i+1],t,k} = \begin{cases} k \in \rho(r\tau, f_0, b): & (1-\delta)\hat{c}_{[d,i],r\tau,k}, \\ \text{otherwise:} & \hat{c}_{[d,i],r\tau,k}, \end{cases} \tag{12}$$

where

$$\delta = \frac{\alpha_{[d,i+1],b,r\tau}}{\sum_{k \in \rho(r\tau, f_0, b)} \hat{c}_{[d,i],r\tau,k}}. \tag{13}$$

If (11) is false, it means that the recording no longer contains the amplitude required by the current note signal. In this case all the remaining amplitudes are transferred to the current note. This leaves no amplitude on the respective bins,

$$\hat{c}_{[d,i+1],r\tau,k} = \begin{cases} k \in \rho(r\tau, f_0, b): & 0, \\ \text{otherwise:} & \hat{c}_{[d,i],r\tau,k}. \end{cases} \tag{14}$$

The current amplitude spectrum of the notes to be isolated can be calculated in step $[d, i]$ as

$$\hat{\boldsymbol{s}}_{i,[d+1],r\tau} = \hat{\boldsymbol{s}}_{i,[d],r\tau} + \left(\hat{\boldsymbol{c}}_{[d,i-1],r\tau} - \hat{\boldsymbol{c}}_{[d,i],r\tau}\right), \tag{15}$$

which is the starting value of the amplitude spectrum of the separated note signals in the next step $[d+1]$. The starting value of the remaining amplitude spectrum in step $[d+1]$ is

$$\hat{\boldsymbol{c}}_{[d+1,0],r\tau} = \hat{\boldsymbol{c}}_{[d,I],r\tau}. \tag{16}$$

The separation finishes after step $[D-1]$. The separated instruments are represented by $\hat{\boldsymbol{s}}_{[D,i],r\tau}$, and the amplitude spectrum of the noise component $\boldsymbol{d}$ is

$$d_{k,t} = \hat{\boldsymbol{c}}_{[D,0],r\tau,k}. \tag{17}$$

For transforming the signal back to time domain the phase information is also needed. This is available from (5) as the separation preserves the phases of the input mixture.

7. Beating correction and conversion back to time domain

When two notes have overtones that overlap in frequency, their mixture will suffer from cancellation effects. If the two frequencies are close to each other but not equal, the

resultant amplitude on these frequencies will fluctuate between the sum and the difference of the two amplitudes. The properties of the fluctuation depend on the phase difference and frequency deviation of the two waves. This phenomenon is called *beating*, and it is generally unnoticeable in a recording due to the complex structure of the polyphonic mixture. As we do not perform a perfect reconstruction of the note signals (we neglect phase relations in the separation algorithm), the beating is not resolved and is still present in the separated signals. However, while it was unnoticeable in the original recording, it is audible when a separated note signal is played alone.

The first effect can be overcome by applying a postprocessing step to the output signals. The amplitude envelope of the signals can be aligned to that of the parent sample by amplifying the separated note signal where its amplitude envelope does not follow the envelope of its parent sample. This postprocessing step eliminates most of the beating from the separated note.

Experiments were carried out to validate the effect of the beating correction step on human listeners. Test subjects were presented separated note signals with and without the beating correction applied. In most cases the listeners reported that the separated notes sounded more natural with beating correction than without it.

However, we must keep in mind that this correction modifies the energy content of the notes. After this step it is not possible to get the original signal back by linearly remixing the separated notes into one channel.

After the beating correction step the notes and the remaining part are reconstructed via inverse Frequency Estimation and inverse STFT.

It is important to mention that while it is beneficial to apply beating correction to the notes we want to isolate, other notes that are not subject to separation should be transformed back to time domain unchanged. This ensures that the notes of interest will be of high quality, even with somewhat modified energy content; while the other notes that are remixed back to the recording do not generate new cancellation effects.

8. Synthetic tests

The quality of the separation system described so far has been evaluated by using a number of synthetic tests. This section features different test scenarios that illustrate the quality of the given method.

The test setup used for evaluating the performance of the separation system is shown in Figure 3. The test system was based upon the instrument sample collection of the University of Iowa [43]. The waveforms were normalized and converted to mono, 44100 Hz samples/sec, 16 bits. Their DC offset was also corrected (shifted to zero) where it was necessary. The waveforms were then divided into samples containing only one instrument note. Samples shorter than 500 ms were dropped, while samples longer than 2 seconds were cropped to 2 seconds. The above described process resulted in 3841 waveforms of

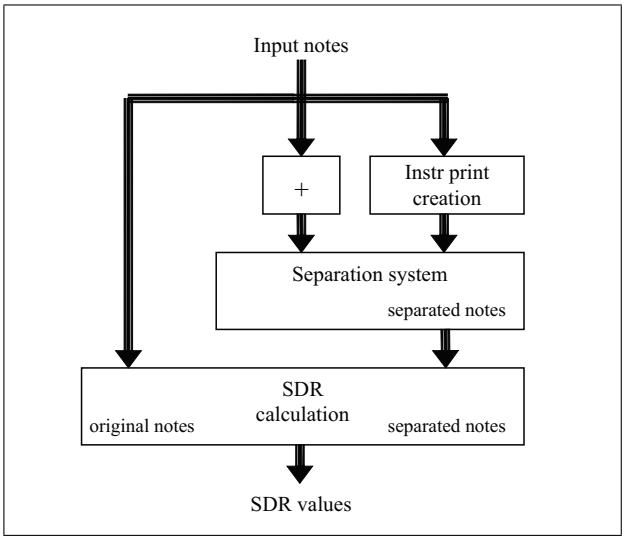

Figure 3. Block diagram of the automatic synthetic test system.

separate instrument notes of harmonic instruments. The instrument database contained samples of the following instruments: flute, saxophone, bass, clarinet, bassoon, trombone, cello, horn, oboe, piano, pizzicato, trombone, trumpet, tuba, viola and violin.

In each of our tests a random set of instrument note waveforms were selected. The waveforms were converted to instrument prints using the technique described in section 3. The selected samples were then mixed together and fed to the separation system as the input recording.

Four series of tests were performed, each one focusing on a specific aspect of note separation. In each series the performance of the system was tested for polyphony levels of 2 to 10. At each level a total of 50 individual tests were carried out. The input and output signals were compared and the error levels were measured.

In the first series the input waveforms were simply mixed together as in

$$\tilde{c}(n) = \sum_{i=1}^{I} \tilde{s}_i^{\text{orig}}(n), \tag{18}$$

where $\tilde{c}$ and $\tilde{s}_i^{\text{orig}}$ are the waveform of the mixed signal, and the note signals, respectively. This method ensured that all the notes shared the same onset. As the onset/offset and note fundamental frequencies were all known, user input was not necessary. Then the separation step was performed with the input mixed signal, the score data and instrument prints that were created previously from the sample database.

The output note waveforms of the system were then compared to the original ones. Since the algorithm preserves the phase information of the input mixture, an error between the original and the output waveforms could be obtained in most cases by subtracting the output signal from the original ones in time domain. The mean-square level of each error signal was computed over the whole

signal. A signal-to-distortion ratio could then be obtained by comparing these levels to the original,

$$\mathrm{SDR}_i = 10\log_{10}\frac{\sum_n \tilde{s}_i^{\mathrm{orig}}(n)^2}{\sum_n \left[\tilde{s}_i(n) - \tilde{s}_i^{\mathrm{orig}}(n)\right]^2} \tag{19}$$

where $\tilde{s}_i$ is the waveform of the separated signal for note i. Within one polyphony level the average and standard deviations of the individual SDR_i values were calculated. This is shown in Figure 4.

When listening to the separated output channels, it was found that the SDR gave incorrect values in some cases. Although the quality of certain separated notes was satisfactory for the human ear, the calculated SDR value still indicated very low separation quality for these notes. After investigating the issue we found that in most of these cases the bad SDR scores can be explained by phase distortion between the original signals and the mixture that was generated from them. When two or more of the original signals contain energy at common frequencies then the phase of the resulting mixture may deviate from the phase of the original signals. This difference is hardly audible to human listeners; however, it highly affects the SDR value.

To overcome this issue we must realize that under normal circumstances a human listener is not able to sense any difference between two signals that differ by phase, but otherwise have the same energy content (except for very rare cases or some synthetic tests). Therefore we introduce here another measure that is based on human audio perception. Instead of comparing the time-domain waveforms to the time-domain residual, we recalculate the SDR error levels using the frequency-domain representation of the signals. The error can now be expressed as

$$\mathrm{SDR}_i^F = 10\log_{10}\frac{\sum_{\forall r\tau}\sum_{k=0}^{K} {s_{i,r\tau,k}^{\mathrm{orig}}}^2}{\sum_{\forall r\tau}\sum_{k=0}^{K}\left[s_{i,r\tau,k} - s_{i,r\tau,k}^{\mathrm{orig}}\right]^2}. \tag{20}$$

Although the new measure represents the perceptual separation quality more efficiently, in this paper we still include the original SDR measurements for easier comparison with other works. We must note, that there is no definite relation between the two.

The second test series was very similar to the first one. In this case, however, the note onset times were dilated by a Δt_i random number, as in

$$\tilde{c}(n) = \sum_{i=1}^{I} \tilde{s}_i^{\mathrm{orig}}(n - \Delta t_i S), \tag{21}$$

where S is the sampling rate and

$$0\,\mathrm{sec} < \Delta t_i < 1\,\mathrm{sec}. \tag{22}$$

The first test provided greater overlap of the notes in time, while in the second test it was easier to observe artifacts in one separated note caused by the onset of another note. However, the two tests returned similar SDR values. Measurement results are shown in Figure 5.

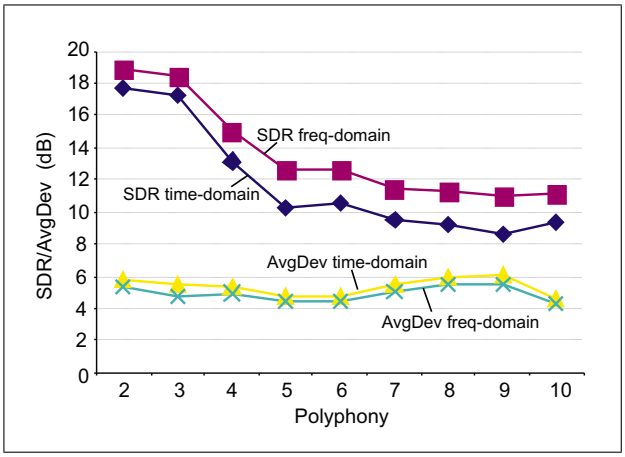

Figure 4. Signal to Distortion Ratio for Test 1.

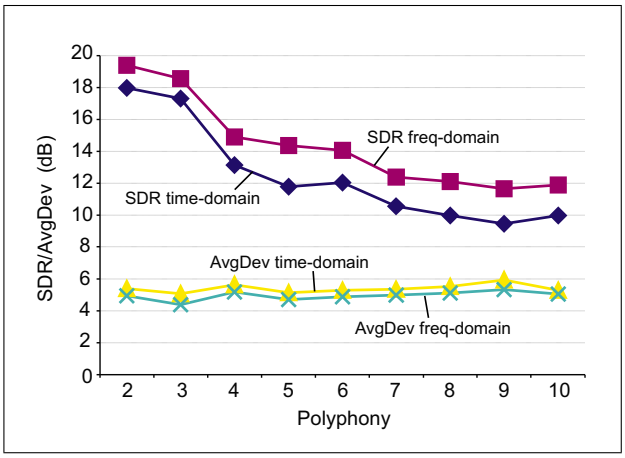

Figure 5. Signal to Distortion Ratio for Test 2.

The third test covered cases with notes that are in close overtone relation with each other. For this reason Test 1 was repeated with input channels that meet the following condition:

$$\forall i \quad \Rightarrow \quad \exists j \neq i \quad \ni \qquad \frac{f_{0,i}}{f_{0,j}} \approx \frac{m}{n}, \quad m,n \in \{1,2,3,4,5\}, \tag{23}$$

where i and j represent the input note signals. Condition (23) ensures that each input notes is in close overtone relation with at least one other input note in the set.

Figure 6 presents the measurement results for Test 3. As expected, the SDR values in this test were lower than in the previous series. This can be attributed to the fact that the vibrations generated by the notes cancelled each other in most of the test cases. Such cases inevitably introduce the beating effect in the generated mixture. The approach we took by applying simplification (5) to the system cannot perfectly handle the beating, thereby leaving some artifacts in the output channels. However, despite the fact that we cannot get back the original tracks, it was found that the separated channels still resemble real-life instruments even at higher polyphony levels.

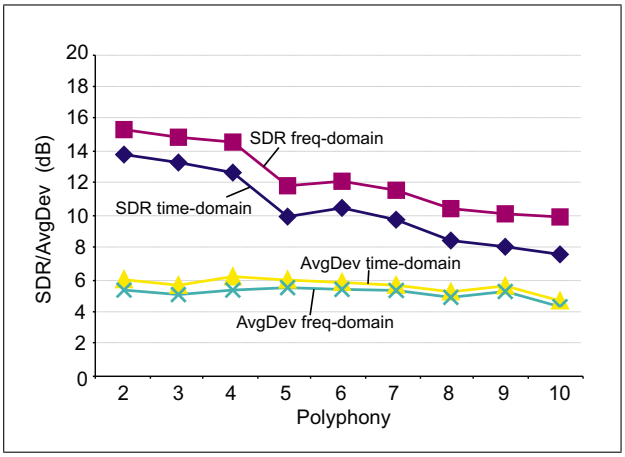

Figure 6. Signal to Distortion Ratio for Test 3.

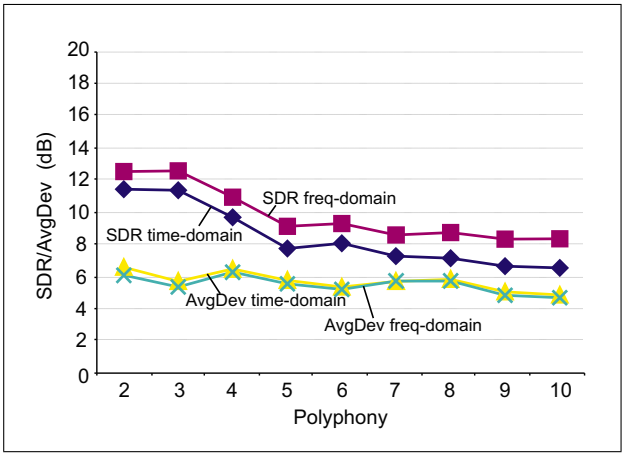

Figure 7. Signal to Distortion Ratio for Test 4.

The last test series focused on cases where the right instrument print is not available. The procedure of the first series was repeated, however, instrument prints were not taken from the input channels. Another sample from the set was used, which was taken from the same type of instrument, but not necessarily from the very same one.

Figure 7 shows the measurement results for Test 4. The results show that the separation quality somewhat dropped when incorrect instrument prints were used. However, such a performance drop is not surprising, given the algorithm is based on good quality prints. Apart from some cases where free intonation instruments (like saxophone) are considered, a deviation between the input channels and the selected instrument prints is typically smaller than the difference of this test. The test results show that the achieved separation quality depends mostly on two factors:

- *Polyphony level:* As this factor increases, the separation quality gradually gets lower. In the background, overtone relations between the separate notes are responsible for less successful separation results. As the polyphony level gets higher, more and more instruments get located at each other's fundamental or overtone frequency. However, this is not surprising; it is all in accordance with human hearing. While we are able to 'hear out' the tune of a violin from a quartet, we may be incapable of performing the same with a full orchestral piece.
- *Quality of instrument prints:* The importance of good prints is revealed in Test 4. Real-life experiments proved also that in many cases it is sufficient to work with prints from the same kind of instrument, while in almost all cases sampling notes from the same instrument provides good quality separation output.

9. Comparison to alternative techniques

Besides the synthetic results the reader may also be interested in how well the proposed methods perform compared to other approaches. Unfortunately, such a comparison is not easy to make as different researches in the topic have different focus areas. For example, ICA is mainly used in environments where multiple observations of a single sound event are available, while the method described here focuses on single channel sources.

Perhaps the most direct (and fair) comparison can be made with methods based on sinusoidal modeling. Virtanen *et al.* [13] studied methods from this area in their work which mainly differed in the way of handling overlapping partials. In our work we repeated their test procedure combined with our separation method. We used the same set of test samples [43] and tested our algorithm on the same polyphony levels. We also used the same method for measuring SDR, although, as earlier stated, this measure provides worse results for our method than the more appropriate SDR^F measure. We have achieved 18/17/13/10 dB SDR separation quality for 2/3/4/5 concurrent notes (see Figure 4) – as opposed to the 15/10/9/8 dB reported by Virtanen.

Another interesting comparison can be made with NMF methods. Reference [13] gives a good overview on these methods too and also provides measurement results of NMF separation cases. In those measurements the best result for NMF source separation was 9.1 dB SDR^F, calculated as an average of several measurements including polyphony levels of 5, 10, 15 and 20. It is important to note that the number of the musical notes as well as their frequencies are detected automatically by NMF, which occasionally leaves some notes undetected and not separated at all. Obviously, these notes cannot and were not considered in the SDR^F measurements, but their ratio in percentage is given instead (error rate ∼25%). In contrast to NMF, the method in this paper requires manual input of the musical score, which may be a slow, but more reliable process. Also, in our automated test environment the properties of the test notes are known, hence all the input source note signals do have a sibling output signal which can be considered in the final SDR^F measurement. That said, our SDR^F results for similar test conditions showed a quality of 9.05 dB, which is on par with NMF techniques. This is a very good result, especially if we take into consideration

the fact that no notes are excluded from our SDR^F measurement, as were the case with NMF. A perhaps somewhat unfair but theoretically interesting comparison can be made if we too exclude the worst 25% from our output signals, which yields an average SDR^F of 11.1 dB: 2 dB higher than in the NMF case. Of course the quality degrades if the used instrument prints are not too close matches of the notes to be separated.

However, the key point of the algorithm and system proposed here is the ability to separate note signals even when they share the same fundamental frequency. The isolation of such notes is rarely covered by other approaches, while it fits naturally into the proposed method and requires no special attention.

10. Conclusion

We have developed a method for separating the signal of single instrument notes from a recording using prerecorded instrument prints. The global system architecture of the separation process was given, along with the description of its building blocks. We have established a simple, yet powerful model for storing instrument prints, and the Simplified Energy Splitter was proposed as an algorithm for solving the energy redistribution problem.

We have demonstrated the potential of the system on synthetic and real-life test cases. Simulation experiments on generated mixtures of pitched real-life musical instruments were carried out. In these experiments we obtained an average SDR^F above 19 dB for two simultaneous sources, and the quality decreased gradually as the level of polyphony increased. Example waveforms of our synthetic tests as well as real-life cases can be downloaded from http://avalon.aut.bme.hu/ aczelkri/separation.

For recordings that only contain harmonically unrelated notes the algorithm provides very clear results. In real life, however, consonant notes with overlapping overtones are usually favored over dissonant ones. Our test results show that even in cases where some notes are located at each other's fundamental or overtone frequencies the separation provides reasonably good results.

Our system was also used in the production of commercial music CDs. Among others, a part of 'Alexander Scriabin: Étrangeté' (BMC Records, 2005) [44] was digitally retouched using our sound separation software. Track 28, 'Vers la flame', is a fast paced post romantic piano piece containing very fast melody motions and constant use of the sustain pedal that lets a large number of played notes sound concurrently. In this track a set of misplayed notes in a complex mixture were separated and transposed to the right frequency. No obvious traces of manipulation are audible in the modified signal, and the disc itself received very high ratings in several reviews: GRAMOPHONE (08/2006), MUSICA (03/2006), PIZZICATO (11/2005) – just to mention a few.

Acknowledgement

This work has been supported by the fund of the Hungarian Research Fund (grant number T68370)

References

[1] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, B. Ong: Melody transcription from music audio: Approaches and evaluation. IEEE Trans. Audio, Speech, Lang. Proc. **15** (2007) 1247–1256.

[2] H. Thornburg, R. J. Leistikow, J. Berger: Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data. IEEE Trans. Audio, Speech, Lang. Proc. **15** (2007) 1257–1272.

[3] D. F. Rosenthal, E. H. G. Okuno: Computational auditory scene analysis. Lawrence Erlbaum, Mahwah, NJ, 1998.

[4] A. A. Wieczorkowska, A. Czyżewski: Rough set based automatic classification of musical instrument sounds. Electronic Notes in Theoretical Computer Science **82** (2003) 298–309.

[5] J. C. Brown: Computer identification of musical instruments using pattern recognition with cepstral coeffcients as features. J. Acoust. Soc. Am. **105** (1999) 1933–1940.

[6] P. Herrera-Boyer, G. Peeters, S. Dubnov: Automatic classification of musical instrument sounds. J. New Music Res. **32** (2003) 3–21.

[7] S. V. Vaseghi: Advanced digital signal processing and noise reduction, second edition. John Wiley and Sons, 2000. 333-354.

[8] I. Kauppinen, K. Roth: Improved noise reduction in audio signals using spectral resolution enhancement with time-domain signal extrapolation. IEEE Trans. Speech, Audio Proc. **13** (2005) 1210–1216.

[9] M. R. Every, J. E. Szymanski: Separation of synchronous pitched notes by spectral filtering of harmonics. IEEE Transactions on Audio, Speech, and Language Processing **14** (2006) 1845–1856.

[10] T. F. Quatieri, R. G. Danisewicz: An approach to co-channel talker interference suppression using a sinusoidal model for speech. IEEE Trans. Acoustics, Speech, and Signal Proc. **38** (1990) 56–69.

[11] M. Davy, S. Godsill: Bayesian harmonic models for musical signal analysis. Proc. of Seventh Valencia International meeting Bayesian statistics, Tenerife, Spain, 2002.

[12] C. Févotte, S. J. Godsill: A Bayesian approach for blind separation of sparse sources. IEEE Trans. Speech, Audio Proc. **14** (2006) 2174–2188.

[13] T. Virtanen: Sound source separation in monoaural music signals. PhD thesis, University of Kuopio, 2006.

[14] T. Virtanen, A. Klapuri: Separation of harmonic sounds using multipitch analysis and iterative parameter estimation. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, USA, 2001, 83–86.

[15] H. Viste, G. Evangelista: A method for separation of overlapping partials based on similarity of temporal envelopes in multichannel mixtures. IEEE Trans. Audio, Speech, Lang. Proc. **14** (2006) 1051–1061.

[16] N. Mitianoudis, M. E. Davies: Using beamforming in the audio source separation problem. 7th Int. Symp. on Signal Processing and its Applications, Paris, July 2003, 89–92.

[17] P. Comon: Independent component analysis, a new concept. Signal Processing, Elsevier **36** (1994) 287–314.

[18] T. Pohle, P. Knees, M. Schedl, G. Widmer: Independent component analysis for music similarity computation.

Proc. 7th International Conference on Music Information Retrieval, Victoria, Canada, 2006, 228–233.

[19] T. Virtanen: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech, Lang. Proc. **15** (2007) 1066–1074.

[20] D. D. Lee, H. S. Seung: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–791.

[21] P. Smaragdis, J. C. Brown: Non-negative matrix factorization for polyphonic music transcription. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, USA, 2003, 177–180.

[22] D. D. Lee, H. S. Seung: Unsupervised learning by convex and conic coding. Proc. Conference on Neural Information Processing Systems **9** (1997) 515–521.

[23] M. Plumbley, S. Abdallah, T. Blumensath, M. Davies: Sparse representations of polyphonic music. EURASIP Signal Processing Journal **86** (2006) 417–431.

[24] T. Virtanen: Sound source separation using sparse coding with temporal continuity objective. Proc. Int. Computer Music Conf., Singapore, 2003.

[25] D. Barry, R. Lawlor, E. Coyle: Sound source separation: Azimuth discrimination and resynthesis. Proc. 7th Int. Conf. on Digital Audio Effects, Naples, Italy, 2004.

[26] J. C. Brown, M. S. Puckett: A high resolution fundamental frequency determination based on phase changes of the fourier transform. J. Acoust. Soc. Am. **94** (1993) 662–667.

[27] K. Aczel, I. Vajk: Instrument prints in note separation of polyphonic music. Proc. WSEAS International Conf. on Signal Proc., Robotics and Automation, Cambridge, England, 2008.

[28] S. McAdams: Recognition of auditory sound sources and events. Thinking in sound: The cognitive psychology of human audition. Oxford University Press, 1993.

[29] S. Handel: Timbre perception and auditory object identification. – In: Hearing. B. C. J. Moore (ed.). Academic Press, New York, 425–463.

[30] E. B. George, M. J. T. Smith: Analysis-by-synthesis / overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. J. Audio Eng. Soc. **40** (1992) 497–516.

[31] K. Vos, R. Vafin, R. Heusdens, W. Kleijn: High-quality consistent analysis-synthesis in sinusoidal coding. Proc. 17th Audio Eng. Soc. Int. Conf., 1999, 244–250.

[32] S. Mallat, Z. Zhang: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Proc. **41** (1993) 3397–3415.

[33] R. Heusdens, R. Vafin, W. B. Kleijn: Sinusoidal modeling using psychoacoustic-adaptive matching pursuits. IEEE Signal Proc. Letters **9** (2002) 262–265.

[34] M. Goodwin, M. Vetterli: Matching pursuit and atomic signal models based on recursive filter banks. IEEE Trans. Signal Processing **47** (1999) 1890–1902.

[35] L. Daudet: Sparse and structured decompositions of signals with the molecular matching pursuit. IEEE Trans. Audio, Speech, Lang. Proc. **14** (2006) 1808–1816.

[36] A. Bultan: A four-parameter atomic decomposition of chirplets. IEEE Trans. Signal Proc. **47** (1999) 731–745.

[37] R. Gribonval: Fast matching pursuit with a multiscale dictionary of Gaussian chirps. IEEE Trans. Signal Proc. **49** (2001) 994–1001.

[38] P. Leveau, E. Vincent, G. Richard, L. Daudet: Instrument-specific harmonic atoms for mid-level music representation. IEEE Trans. Audio, Speech, Lang. Proc. **16** (2008) 116–128.

[39] P. Iverson, C. L. Krumhansl: Isolating the dynamic attributes of musical timbre. J. Acoust. Soc. Am. **94** (1993) 2595–2603.

[40] E. Zwicker, G. Flottorp, S. S. Stevens: Critical band width in loudness summation. J. Acoust. Soc. Am. **29** (1957) 548–557.

[41] S. W. Smith: The scientist and engineer's guide to digital signal processing. California Technical Publishing, 1997.

[42] B. Edler, H. Purnhagen: Parametric audio coding. Proc. IEEE Int.Conf. on Communication Technology, 1, 2000, 614–617.

[43] The University of Iowa musical instrument samples database. (02/08/2010): `http://theremin.music.uiowa.edu/MIS.html`.

[44] Gábor Csalog plays "Alexander Szkrjabin: Étrangeté". BMC Records, 2005, (02/08/2010): `bmcrecords.hu/pages/frameset/direkt.php?kod=099`.