Simple and powerful instrument model for the source separation of polyphonic music

KRISTÓF ACZÉL[1], ISTVÁN VAJK[2]
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
3-9. Muegyetem rkp. , H-1111, Budapest
HUNGARY
aczelkri@aut.bme.hu[1], vajk@aut.bme.hu[2]

Abstract: - This article presents a new approach to sound source separation. The introduced algorithm is based on spectral modeling of real instruments. The separation of independent sources is carried out by dividing the energy of the mixture signal based on these instrument models. This way it is possible to regain some of the information that was lost when the independent sources were mixed together into a single signal. The paper presents the theory behind the proposed separation system, then focuses on the instrument model that is the basic element of the approach. Measurement results are given for polyphony levels from 2 to 10 demonstrating the separation quality, with special regard to the effect of prints on the result.

Key-Words: - sound separation, instrument print, polyphonic music, energy split

1 Introduction

The separation of instrument tracks in a polyphonic music piece has been a considerable challenge for researchers for a very long time. The possibility of correcting or altering existing musical recordings would open a whole new perspective in sound processing. Isolation of musical notes could not only allow filtering out and fixing bad notes in a recording, but also altering the polyphonic structure in other ways like pitch shifting, volume controling, formant adjusting etc.

The complexity of separation can be attributed to the fact that the information to be retrieved is actually not present in the signal. Since a regular listener is not interested in the separate tracks, only a few (typically one or two) channels are used for storing the recording. Therefore more instruments are usually downmixed into the same channels. There are several different approaches for regaining the information that is lost in this step.

[1] introduces a sound source separation algorithm that requires no prior knowledge on the instrument notes in the recording, and performs the task of separation based purely on azimuth discrimination within the stereo field. Although results are impressive, separating individual notes is out of the focus, only instrument groups are considered.

[3], [4], [5] describe a method which separates harmonic sounds by applying linear models for the overtone series of the sound. 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 [2].

Beamforming techniques [6], [7] along with the Independent Component Analysis framework offer a different way of separation. A relatively large array of microphones is employed in the time the recording is made. The travel time of the time signal and the difference of the numerous recorded signals is used in calculations that increase the receiver sensitivity in the direction of wanted signals and decrease the sensitivity in other directions. However, these methods rely on certain preliminary conditions and studio setup to achieve good results.

As there are several other approaches to sound source separation, (like Non-Negative Matrix Factorization [10], [11], sparse coding [8], [9], etc.), this article does not aim to provide a complete list of these methods. [12] Provides a good overview of the most common methods and ‘state of the art’.

Previous results of current research have been presented in [16] and [17]. The key element of our approach is to use reference samples of real instruments, called instrument prints. Based on these prints the spectral energy of the original recording is split between the notes to be separated.

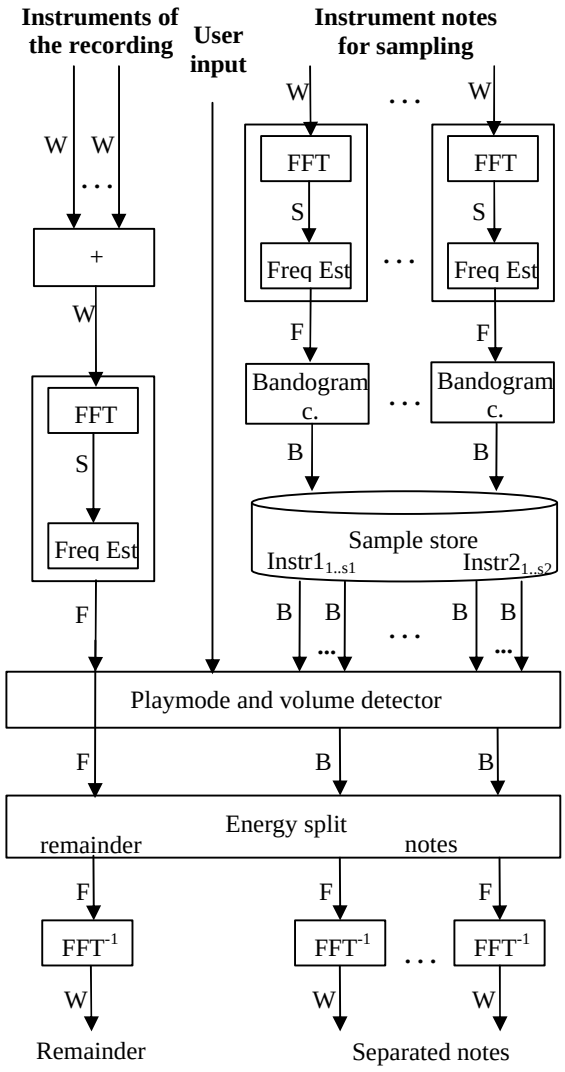

Figure 1: Signal flow and block diagram of the separation process

The following sections will give a brief overview of the separation process, then discusses instrument prints in detail.

2 Overview of the separation process

This section gives an overview of the separation process. Building blocks are described so that the reader gets familiar with the approach. Figure 1 shows the block-diagram of the sound separation process.

Figure 1 uses the following notations for different representations of signals:

- **Simple waveform (W):** time domain function of the signal
- **Simple FFT (S):** simple spectrogram storing the amplitude $c_{k,t}$ and phase $j_{k,t}$ for each bin.
- **Frequency estimated spectrogram (F):** $c_{k,t}$ amplitudes and $j_{k,t}$ phases are the same as for the simple FFT, but an $f_{k,t}^{true}$ true frequency value is stored additionally for each bin.
- **Bandogram (B):** 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 is available either.

First, all input signals are converted from time domain to frequency domain. The way this transformation is carried out is described in Section 2.1. Before any separation can take place, instrument prints have to be generated from instrument note signals. This is covered in Section 2.2 briefly, then elaborated in Section 3. The score of the input music is entered by the user, except for the volume and intonation of the instruments, which is estimated algorithmically (Section 2.3). The method proposed in this paper is based on a certain simplification of the original separation problem. This is covered in Section 2.4 along with the proposed solution for isolating the musical note signals from each other. Finally the separated note signals are converted back to time domain.

The following subsections describe each block in Figure 1 in detail.

2.1 Conversion between time and frequency domain

This section shows the algorithm used to the frequency-domain transformation. The algorithm generates a spectrogram of the recording that is much more precise for musical analysis than the conventional FFT spectrogram.

Earlier literature [13], [14] covered different transformation methods in order to determine the best possible means for the analysis of digitized audio signals. Current research has examined the analysis of digitized polyphonic musical signals in particular. Experiments have been carried out in order to find and validate the appropriate parameters that are the most suitable for musical signals, such as window length, zero-padding ratio etc. These experiments concluded that a frame length of 2048 samples should be used, with an overlap ratio above 1/32. Blackman window has been found to be one of the most effective, yet easy-to-implement window functions that can be applied to the frames

prior to the transformation. Due to its property that it provides energy content that is almost independent from the source frequency in the signal it fits very well into the concept of current research, where our instrument samples store the sum energies of frequency bands. After windowing the signal zero padding of 50% or greater should be used (meaning the 2048-sample-long signal is padded to the length of 4096 or more samples with zeros). A signal prepared in such a way can finally be converted to frequency domain using standard FFT algorithm.

In [15] and [18] a frequency estimation method was introduced, that calculates true frequencies present in the original signal from subsequent phase values. For a frame starting at time *t* the FFT coefficients and phases are $c_{k,t}$ and $f_{k,t}$, respectively. In this paper the time index will be omitted in some of the equations for better understanding. Two subsequent frames are needed by the algorithm for the calculation. Assuming that the frame starts at t_1 and ends at t_2 , a true frequency f_{k,t_2}^{true} can be computed for each bin as follows. Let the frequency of the kth bin be

$$f_k = k\frac{samplerate}{framesize}. \tag{1}$$

The true frequency of each bin will deviate from this value as in

$$f_{k,t_2}^{true} = f_k + \frac{j_{k,t_2}^{dev}}{2p\cdot(t_2 - t_1)}, \tag{2}$$

where

$$\begin{aligned} j_{k,t_2}^{expt} &= j_{k,t_1} + (t_2 - t_1)\cdot 2p\, f_k \\ j_{k,t_2}^{dev} &= j_{k,t_2} - j_{k,t_2}^{expt} + l\cdot 2p \end{aligned}, \tag{3}$$

where $f_{k,t}$ is the phase of bin *k* in time *t*; j_{k,t_2}^{expt} is the expected phase; j_{k,t_2}^{dev} is the deviance between the expected and measured phase; f_{k,t_2}^{true} is the estimated true frequency of bin *k* in time *t* and $l \in Z : -p < j_{k,t_2}^{dev} \le +p$.

Even more precise results can be achieved if the deviance is calculated from the weighted sum of previous and future phase values as in

$$\hat{j}_{k,t_x}^{dev} = \frac{\sum_{x=-\infty}^{\infty} j_{k,t_x}^{dev}\cdot c_{k,t_x}\cdot J(x)}{\sum_{x=-\infty}^{\infty} c_{k,t_x}\cdot J(x)}, \tag{4}$$

where $J(x)$ denotes an arbitrary function, whose typical properties are as follows:

$$J(x_1) < J(x_2) \quad \begin{matrix} \forall\ x_1 < x_2 < 0 \\ \text{where } x_1 x_2 > 0 \end{matrix} \tag{5}$$

The true frequency can be now calculated as:

$$\hat{f}_{k,t_2}^{true} = f_k + \frac{\hat{j}_{k,t_2}^{dev}}{2p\cdot(t_2 - t_1)}, \tag{6}$$

The original estimation algorithm (2) will be referred to as Frequency Estimation (FE), while extension (6) will be called Phase Memory (PM). Figure 2 shows the effect of both algorithms using a real-life music signal.

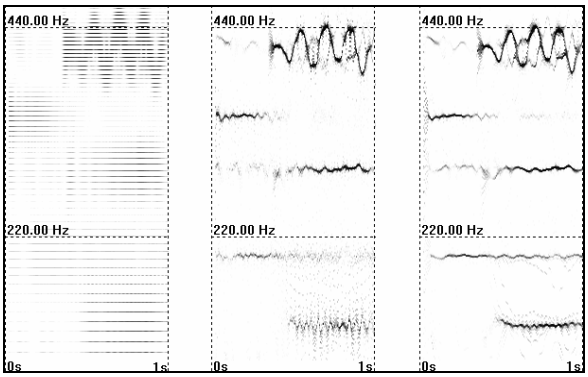

Figure 2: Spectrogram plots. a) simle STFT, b) Frequency Estimation, c) Phase Memory

Instrument print creation, as well as signal analysis and the actual separation step are all based on the methods described above. However, it is important to mention that although this method is a very effective tool of signal analysis, it is not used later in the transformation back to time domain.

2.2 Bandogram and instrument prints

In the bandogram calculation step the frequency-domain signal of real-life instrument notes is converted to instrument samples that are stored in a database. A collection of instrument samples on different base frequencies and with different intonations will be considered an instrument print. The structure of instrument prints will be described in Section 3 in detail.

2.3 Playmode and volume estimator.

While the user can specify the location of the instrument notes in frequency and time, they may

not be capable of deciding the precise intonation and volume level of notes. However, to carry out the energy split step an optimal playmode matrix **M** must be found. For the sake of convenience the volume is also incorporated in **M** from now on.

M is by definition perfect if the separated notes are the perfect replicas of the parent instrument samples that were used in the energy split, and the remaining part is zero. In general, matrix **M** is considered good if an energy split step that uses **M** generates notes that 'resemble' their parent sample. The energy split step is carried out with all possible combinations of **M** matrices, and the one causing the least separation error is considered to be the best solution for the separation. Depending on the size and possible values of **M,** the number of steps needed for finding the best combination of the instrument samples may require huge computational power. If we consider the playmode space to be continuous, it is not even possible to iterate through all the combinations. Finding an algorithm faster than brute force iteration, however, is out of the scope of this article.

2.4 Energy split

This section describes the core of the separation process, the energy split. Since the original decomposition problem cannot be solved due to the lack of information, a certain simplification will be proposed that makes it possible to carry out the separation even under these circumstances at the expense of slightly lowered quality. By applying this change the separation problem will be simplified to an energy split problem. After that, the reader will be guided through the implementation of the energy split process itself.

With time being represented as $t = rt$, where r stands for the current frame and t is the time difference between subsequent frames the original separation problem can be drafted as

$$\underline{\mathbf{c}}_{rt} = \sum_{\forall i} \underline{\mathbf{s}}_{i,rt}^{orig} \ , \tag{7}$$

where $\underline{\mathbf{c}}_{rt} = [c_{k,rt} \cdot e^{g_{rt,k}}]$ is the mixed signal which is the input of the separation algorithm and $\underline{\mathbf{s}}_{i,rt}^{orig} = [s_{i,k,rt}^{orig} \cdot e^{s_{i,k,rt}}]$ values represent the original notes. This undetermined system of equations cannot be solved unambiguously without any further constraints.

Our knowledge on the original notes is rather limited, therefore it is not possible to separate the recording to notes that are the perfect replicas of the original ones. As the original separation problem cannot be solved, simplifications have to be made, the most obvious being the elimination of the unknown $\hat{s}_{i,k,0t}$ phases from the equation system:

$$\hat{s}_{i,rt,k} = g_{rt,k} \tag{8}$$

This formulates the original problem as

$$\underline{\mathbf{c}}_{rt} = \sum_{\forall i} \underline{\hat{\mathbf{s}}}_{i,rt} + \underline{\hat{\mathbf{c}}}_{rt} \tag{9}$$

where $\underline{\hat{\mathbf{s}}}_i$ stands for separated note i and $\underline{\hat{\mathbf{c}}}$ is the remaining energy in the recording after the separation. This modification, motivated by the characteristics of human perception, exploits the fact that the human ear can not differentiate by the phase of the heard sinusoids, only hears magnitude differences. The quality impact of this modification is not discussed in this paper, however we note that our experiments have shown this tradeoff to be acceptable in most cases.

In the energy split step bandograms of the right samples are used to recreate spectrograms of the notes to be separated from the remaining part of the recording. Semi-linear decomposition is used for this purpose. The exact frequency, volume and playmode of all notes to be separated are assumed to be known. The iterative algorithm used to divide the energy between the target notes starts out with the original Frequency Estimated FFT spectrogram of the recording. In one step a fraction of the energy of the selected reference samples is transferred from the FFT of the recording to the FFT of the separated notes. This ensures a fair division of the energy of the recording between the notes. Any energy after the last step is considered noise and is not added to any of the separated notes' spectrograms.

3 Instrument prints

The energy split divides the energy in the recording between the notes. In cases where the notes in the recording do not overlap in time or frequency this is a very straightforward task. However, overlapping notes make it necessary to divide the full energy between the notes to be separated. A decision has to be made regarding the ratios of the original energy that will be transferred from the FFT spectrogram of the recording to different separated notes. Instrument prints will help make that decision.

This section deals with the instrument model that is used in the energy split process. First a short overview on the properties of natural instruments will be presented, then the proposed instrument model is covered in detail.

3.1 Instrument note basic features

The most basic signal features of natural instrument notes are their base frequency along with the corresponding subjective attribute pitch and the noise-like component they carry. In other words, most of the pitched sounds are complex waveforms consisting of several components that can be categorized either as a periodic or aperiodic component. Periodic components are called partials or harmonics. The frequency of each such component is the multiple of the lowest frequency f_{base}, called the fundamental frequency. Their time function can be expressed as

$$x(t) = \sum_{p=1}^{P} a_p \cdot \sin(2p \cdot p \cdot f_{base} \cdot t + j_0) . \qquad (10)$$

where a_p is the amplitude of the p^{th} partial, P is the total number of partials and j_0 is the starting phase. The aperiodic component has a noise-like waveform, and its time function cannot be effectively predicted.

The perceptual counterpart of frequency is pitch, which is a subjective quality often described as highness or lowness. Although the pitch of complex tones is usually related to the pitch of the fundamental frequency, it can be influenced by other factors such as for instance timbre. Some studies have shown that one can perceive the pitch of a complex tone even though the frequency component corresponding to the pitch may not be present (denoted as a missing fundamental) [19] (pp 274.) Is it out of the scope of this paper to review the literature dealing with pitch perception.

In western music, the pitch scale is logarithmic, i.e. adding a certain interval corresponds to multiplying a fundamental frequency by a given factor. Then, an interval is defined by a ratio between two fundamental frequencies f_1 and f_2. For an equal-tempered scale, a semitone is defined by a frequency ratio of

$$\frac{f_2}{f_1} = 2^{\frac{1}{12}} . \qquad (11)$$

An interval of n semitones is defined by

$$\frac{f_2}{f_1} = 2^{\frac{n}{12}} , \qquad (12)$$

that is, the interval in semitones n between two fundamental frequencies f_1 and f_2 is defined by

$$n = 12 \cdot \log_2 \left(\frac{f_2}{f_1} \right) \qquad (13)$$

The first harmonic frequencies of a tone with the approximate intervals from the fundamental frequency are represented in Table 1.

In western music notation and equal-tempered scale, fundamental frequencies are quantized to pitch values using a resolution of one semitone. The *A* 440 Hz is considered as the standard reference frequency, although we cannot assume that orchestras are always be tuned to this pitch.

Harmonic	Freq	Approximate interval with f_{base}	Pitch class
1	f_{base}	unison	A
2	$2 \cdot f_{base}$	octave	A
3	$3 \cdot f_{base}$	octave + 5th	E
4	$4 \cdot f_{base}$	2 octaves	A
5	$5 \cdot f_{base}$	2 octaves + major 3rd	C#
6	$6 \cdot f_{base}$	2 octaves + 5th	E
7	$7 \cdot f_{base}$	2 octaves + 7th	G
8	$8 \cdot f_{base}$	3 octaves	A

Table 1: Intervals between the first 8 harmonics of a complex tone and its fundamental frequency f_{base}. Example for the harmonics of A

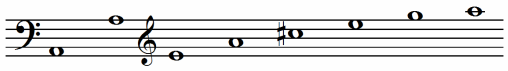

Figure 3: Harmonic series from A2

The importance of understanding the described structure of instrument notes becomes apparent if we consider that in many music cultures harmonic notes are usually favored over inharmonic ones. This means that some of the harmonics of one note are very likely to coincide with the base tone or a harmonic of another one. In these cases the energy splitter needs a hint on what proportion of the full energy on a certain frequency belongs to the different simultaneously playing notes.

3.2 Instrument print structure

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. At some point we will definitely have to feed additional information into the separation system to complete the missing data. Human listeners, who are known to be able to do the separation in their mind, use memories of instruments and memories of the notes in the musical piece being performed. This is their source of additional information. Copying nature has many times been proven to be the right

approach. Based on the findings of the previous section, this section shows a way of implementing a memory of known instruments, trying to mimic the way the human brain works.

[2] describes a method for separating sounds without prior knowledge. The proposed method is powerful as it is based on a simple but general feature of harmonic pitched instruments. It exploits the fact that pitched instruments generate vibrations typically on the base frequency and overtones, and the relationship between the energy levels on these frequencies usually meet the spectral smoothness principle. However, the algorithm in [2] may not be able to separate two notes on the same frequency without any prior information of their properties. For this reason we will store a certain representation of instruments. This representation will be called an instrument print. [20] presents experiments that examine the dynamic attributes of timbre evaluating the role of onsets in similarity judgments. It also gives an overview of researches pursuing the identification of the most important properties of instrument sounds that allow a human listener to distinguish them. The instrument prints in this paper are partly based on these researches, in the sense that they contain the features that were found important in the aforementioned experiments. However, separation purposes require more information on instruments than pure identification does.

An instrument print contains samples from an instrument on different frequencies and with different intonations, '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 dimensions, depending on the way instrument can be played. These cannot always be defined by mathematical definitions, very often they can only be expressed by subjective terms (e.g. loudness, sharpness, warmth etc.). The instrument print is a collection of samples on different frequencies f and also with different values in the playmode space $\underline{\mathbf{M}} = [m_1, m_2, ...m_p]$. It can be regarded as a function

$$\underline{\mathbf{A}}(\underline{\mathbf{M}}, f_k, f_{base}, t) \qquad (14)$$

with the conditions

$$t, m_x, f_{base} \in \mathbb{R}^+$$
$$0 < m_x < m_{x,\max}$$
$$0 \le t < \infty$$
$$0 < f \le 20000Hz$$

which shows how amplitudes change over time over the frequency range for a specific note at a specific frequency f_{base} played with a specific playmode **M**. Figure 4 depicts the FE spectrogram of a simple piano note.

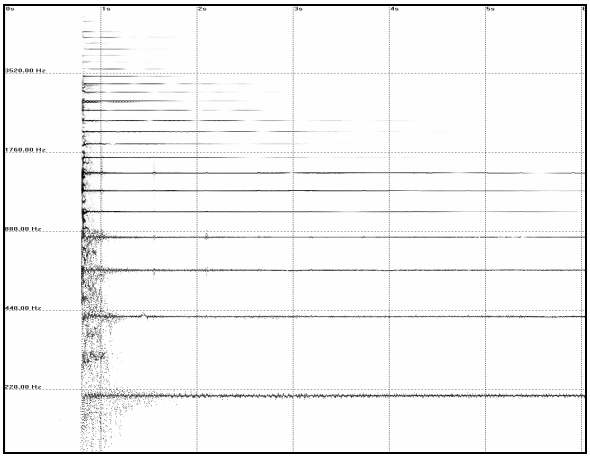

Figure 4: FE Spectrogram of piano note A3

In reality, we will not have all the samples an instrument can produce, only a few of them, on different frequencies and playmodes. Our samples will also be finite in time. Furthermore, a sample will not store a continuous spectogram, only the energy characteristics in certain frequency subbands. This will be called a 'bandogram' (Figure 5). The subbands are aligned on a logarithmical frequency scale. The sum of the energy in the subbands will be calculated and stored in the bandogram. One sample is calculated from a signal that contains exclusively one note originating from the instrument, as in

$$A_{\mathbf{M}, f_{base}, b, rt} = \sum_{f_{base} \cdot 2^{\frac{b-0,5}{R}} < \hat{f}_{k,rt}^{true} < f_{base} \cdot 2^{\frac{b+0,5}{R}}} c_{k,rt} \qquad (15)$$

where

$$b = \left\lfloor \log_{\sqrt[R]{2}} \frac{f_{base}}{\hat{f}_{k,rt}^{true}} \right\rfloor$$

identifies the specific subband, while R is an experimental value defining the resolution in frequency range, that is, the number of subbands per octave. Experiments showed that $R = 12$ provides good enough resolution in log frequency, and it is also easy to understand since an octave consists of 12 semitones.

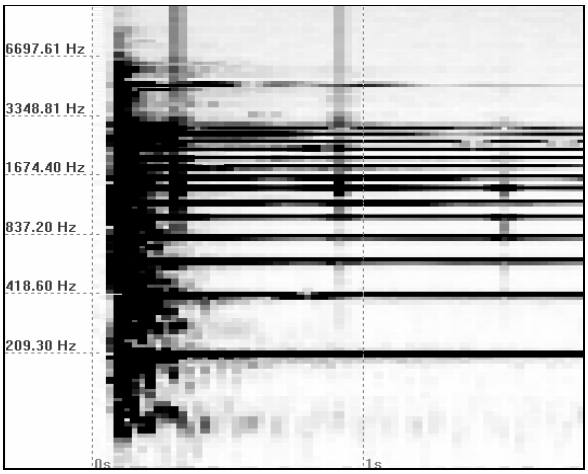

Figure 5: First 2 seconds of an instrument sample (or bandogram) created from piano note A3

As mentioned earlier, the number of recorded instrument samples is finite both in frequency and playmode spaces. However, the energy split step requires samples on virtually any frequency and playmode. When a sample is needed that is not contained by the instrument print, the needed sample must be interpolated from the existing ones. Currently linear interpolation is used, the best interpolation method is subject to future research. If the playmode space is one-dimensional, the interpolation can be written as follows:

$$A(m, f_{base}, b, rt) = A_{m_{-1}, f_{-1}, b, rt} \cdot \frac{f_{+1} - f_{base}}{f_{+1} - f_{-1}} \cdot \frac{m_{+1} - m}{m_{+1} - m_{-1}} + A_{m_{+1}, f_{+1}, b, rt} \cdot \frac{f_{base} - f_{-1}}{f_{+1} - f_{-1}} \cdot \frac{m_{+1} - m}{m_{+1} - m_{-1}} + A_{m_{+1}, f_{+1}, b, rt} \cdot \frac{f_{+1} - f_{base}}{f_{+1} - f_{-1}} \cdot \frac{m - m_{-1}}{m_{+1} - m_{-1}} + A_{m_{+1}, f_{+1}, b, rt} \cdot \frac{f_{base} - f_{-1}}{f_{+1} - f_{-1}} \cdot \frac{m - m_{-1}}{m_{+1} - m_{-1}} \quad (16)$$

where $f, rt \in \mathbb{R}$, $o \in \mathbb{N}$, m_{-1} and m_{+1} are the previous and next closest sampled playmodes, f_{-1} and f_{+1} are the previous and next closest sampled frequencies.

4 Synthetic tests

The quality of the separation system described so far has been evaluated by using a number of synthetic tests. As this article focuses on instrument prints we will present test scenarios here that illustrate the quality of the given method from this perspective.

The test setup used for evaluating the performance of the separation system is shown in Figure 6. The test system was based upon the instrument sample collection of the University of Iowa, [23]. The waveforms were normalized and converted to mono, with a sampling rate of 44000Hz, 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 using tresholding. Samples shorter than 500 ms were dropped, while samples longer than 2 seconds were cropped to 2 seconds. The above mentioned process resulted in 3841 waveforms of 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 strings, 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 3.2. The selected samples were then mixed together and fed to the separation system as the input recording.

Three testing series were performed. 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.

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

$$\tilde{c}(n) = \sum_{i=1}^{I} \tilde{s}_i^{orig}(n), \quad (17)$$

where $\tilde{c}$ and $\tilde{s}_i^{orig}$ are the waveform of the mixed signal, and the note signals, respectively. This method ensures that all the notes share the same onset time. As the onset/ending times and note base frequencies were all known, user input was not necessary. This part was algorithmically fed to the separation system. Then the separation step was performed with the input mixed signal, the score data and the pre-sampled instruments.

The output note waveforms of the system were then compared to the original waveforms. Since the algorithm preserves the phase information of the original (mixed) input signal, an error between the original and the output waveforms can be obtained simply by subtracting the output signal from the original ones in time domain.

Low-level measures are simple statistics of the separated and reference signals. The signal-to-distortion ratio (SDR) has been used many times in literature as a measure to describe the quality. It is the ratio of the energies of the reference signal and the error between the separated and reference signal, defined in decibels. In the separation of music

signals, Jang and Lee [21] reported average SDR of 9.6 dB for an algorithm which trains basis functions separately for each source. Helén and Virtanen [22] reported average SDR of 6.4 dB for their algorithm in the separation of drums and polyphonic harmonic track. Also the terms signal-to-noise or signal-to-residual ratio have often been used to refer to the SDR.

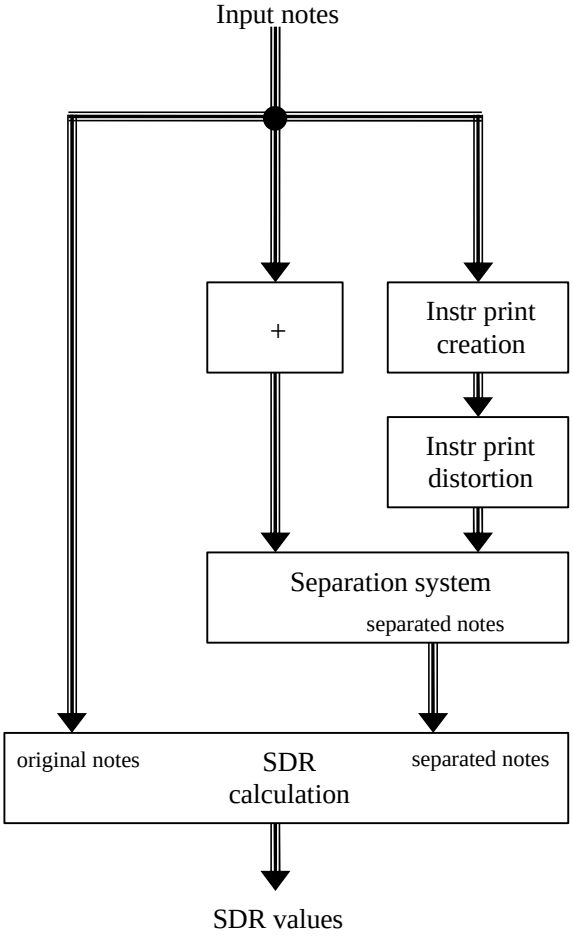

Figure 6: Block diagram of the automatic synthetic test system

In our measurements 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:

$$SDR_i \text{ [dB]} = 10\log_{10}\frac{\sum_n \tilde{s}_i^{orig}(n)^2}{\sum_n \left[\tilde{s}_i(n)-\tilde{s}_i^{orig}(n)\right]^2}, \quad (18)$$

where $\tilde{s}_i$ is the waveform of the separated signal for note *i* for polyphony level *l*.

Within one polyphony level the average and standard deviation of the individual $SDR_{l,i}$ values were calculated, as shown in Figure 7.

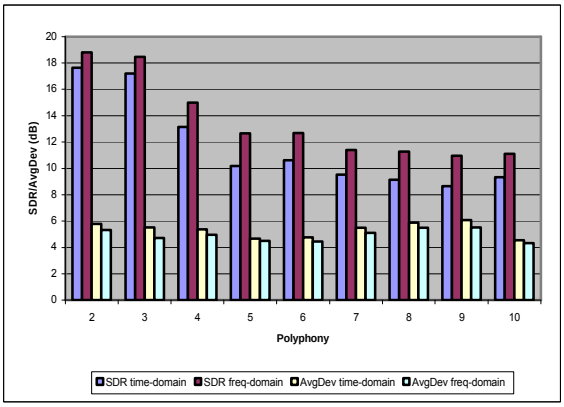

Figure 7: Signal to Distortion Ratio for test 1

The SDR value shows to what extent two signals are different from each other. However, there are some distortions that the human ear is not able to perceive. Two signals with the same energy content but different starting phases will typically be considered identical by human listeners, while the SDR value may indicate huge difference between them. For this reason we propose another, similar measure for testing similarity between signals. The new measure is much like the original SDR, but is calculated in frequency domain instead of time domain. It can be expressed as:

$$SDR_i^F \text{ [dB]} = 10\log_{10}\frac{\sum_{\forall rt}\sum_{k=0}^{K} s_{i,k}^{orig}(rt)^2}{\sum_{\forall rt}\sum_{k=0}^{K}\left[s_{i,k}(rt)-s_{i,k}^{orig}(rt)\right]^2}. \quad (19)$$

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

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

$$\exists i \neq j \ni \quad \frac{f_{base,i}}{f_{base,j}} \approx \frac{m}{n} \quad m,n \in \{1,2,3,4,5\} \quad (20)$$

where *i* and *j* represent the input instrument notes. Condition (20) ensures that any two notes that are mixed together for testing will be in close overtone relation with at least one input note in the set.

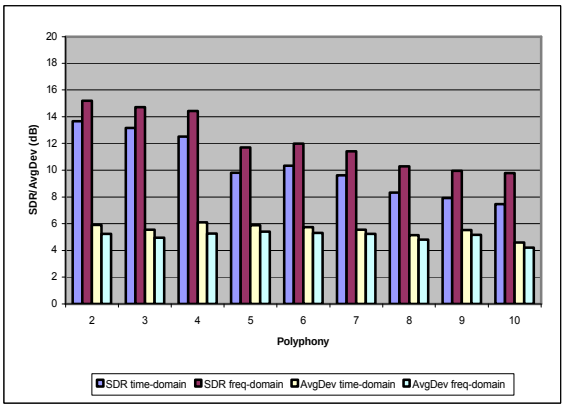

Figure 8: Signal to Distortion Ratio for test 2

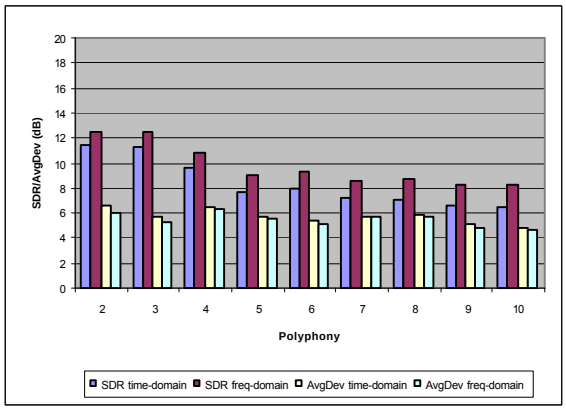

Figure 9: Signal to Distortion Ratio for test 3

Figure 8 presents measurement results for test 2. As can be expected, the SDR and SDR^F 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 (8) to the system does not 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 in higher polyphony levels.

The third 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 9 shows the measurement results for test 4. The results show that the separation quality somewhat dropped when incorrect instrument prints were used. However, this was expectable for an algorithm that bases on the use of 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 in 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 the very cause of less succesful separation results. As the polyphony level increases, more and more instruments get located at each other's base 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 quartett, we may be incapable of doing the same with a full orchestral piece.
- *Quality of instrument prints:* The importance of good prints is revealed in test 3. 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.

5 Conclusion

The paper has introduced a method for separating instrument notes in recordings using pre-recorded instrument prints. A model for storing the properties of real instruments was proposed, and the usability and effectiveness of a system that uses instrument prints was proven. The results are quite promising. For recordings that contain harmonically unrelated notes only, 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 on each other's base or overtone frequencies the separation provides reasonably good results.

The availability of the instruments from the recording ensures good quality instrument prints. In this case we have achieved over 18dB average SDR^F for two sources. For signals with strong harmonic constraints the SDR^F was around 15dB. while in the case of poorer prints the separation quality still reached 12dB.

Examples of some of the synthetic test cases, along with some other samples can be downloaded from http://avalon.aut.bme.hu/~aczelkri/separation.

Acknowledgements

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

References

[1] D. Barry, R. Lawlor, E. Coyle, Sound Source Separation: Azimuth Discrimination and Resynthesis, *Proc. of 7th International Conference on Digital Audio Effects*, DAFX 04, Naples, Italy, 2004.

[2] A. Klapuri, Multipitch estimation and sound source separation by the spectral smoothness principle, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, 2001.

[3] T. Virtanen, A. Klapuri, Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation, *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, 2001.

[4] T. Virtanen, A. Klapuri, Separation of Harmonic Sound Sources Using Sinusoidal Modeling, *Proc of. IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

[5] T. Virtanen, A. Klapuri, Separation of harmonic sounds using linear models for the overtone series, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, Fla, USA, 2002.

[6] N. Mitianoudis, M. E. Davies, Using Beamforming in the audio source separation problem, *7th Int Symp on Signal Processing and its Applications*, Paris, 2003

[7] N. Mitianoudis, M. E. Davies, Using Beamforming in the audio source separation problem, *7th Int Symp on Signal Processing and its Applications*, Paris, 2003

[8] S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by nonnegative sparse coding of power spectra, *Proc of International Conference on Music Information Retrieval*, Barcelona, Spain, 2004

[9] S. A. Abdallah, M. D. Plumbley, Unsupervised analysis of polyphonic music by sparse coding, *IEEE Transactions on Neural Networks*, Vol. 17, No. 1. 2006 pp. 179 – 196

[10] Lee, D.D. and H.S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* 401, pp 788 – 791, 1999

[11] P. Smaragdis and J. C. Brown, Non-Negative Matrix Factorization for polyphonic music transcription, in Proc. 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, USA, 2003

[12] T. Virtanen, Sound Source Separation in Monoaural Music Signals, PhD thesis, University of Kuopio, 2006

[13] R. Pintelon, J. Schoukens, *System Identification, A frequency domain approach*, ISBN 0-7803-6000-1, Wiley-IEEE Press, pp. 33-44, 2001

[14] S. Gade, H. Herlufsen, Use Of Weighting Functions in DFT/FFT Analysis (Part I), *Brüel & Kjær Technical Review*, No. 3., 1987

[15] Judith C. Brown: "A high resolution fundamental frequency determination based on phase changes of the Fourier Transform", *J. Acoust. Soc. Am* Vol. 94 No. 2., pp. 662 – 667, 1993

[16] K. Aczél, Sz. Iváncsy, Sound separation of polyphonic music using instrument prints, *Proc of EUSIPCO 2007*, Poznan, Poland, 2007.

[17] K. Aczél, I. Vajk, Note separation of polyphonic music by energy split, *Proc. of WSEAS International Conference on Signal Processing, Robotics and Automation*, Cambridge, England, 2008

[18] S. M. Bernsee, Pitch Shifting Using the Fourier Transform http://www.bernsee.com/dspdimension.com/html/pshiftstft.html (10-04-2008)

[19] Schmuckler, M. A. (2004). Pitch and pitch structures. In *Neuhoff, J. (Ed.), Ecological Psychoacoustics*, pp 271–315. Elsevier.

[20] P. Iverson; C. L. Krumhansl, Isolating the dynamic attributes of musical timbre, The *J. Acoust. Soc. of America*, Vol. 94, No. 5, pp. 2595 – 2603, 1993

[21] G.-J. Jang and T.-W. Lee. "A maximum likelihood approach to single channel source separation". *Journal of Machine Learning Research*, Vol. 4. No. 7-8, pp. 1365 – 1392, 2003.

[22] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negtive matrix factorization and support vector machine". *In Proc of European Signal Processing Conference*, Turkey, 2005.

[23] The University of Iowa Musical Instrument Samples Database. http://theremin.music.uiowa.edu, (31/03/2008).