Instrument prints in note separation of polyphonic music
KRISTÓF ACZÉL\textsuperscript{1}, ISTVÁN VAJK\textsuperscript{2}
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
3-9. Műegyetem rkp., H-1111, Budapest
HUNGARY

Abstract: - Decomposing a polyphonic musical recording to separate instrument tracks or notes has always been a challenge. Such a signal is the superposition of many separate tracks, and it is theoretically impossible to extract the component tracks without the information that was lost at the superposition. One of the ways of note separation is splitting the energy of the recording based on reference models of instruments. This paper introduces a possible structure for these models, concentrating on how they are built into the separation process.

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

1 Introduction
Separation of polyphonic music into isolated instrument tracks has been challenging researchers for a very long time. The possibility of correcting or altering existing musical recordings would open a whole new window 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 controlling, 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 usual listener is not interested in the separate tracks, only a few (typically two or one) 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.

\[1\] is 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 not in the focus, only instrument groups.

\[3\], \[4\], \[5\] 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 \[6\] 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 many 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.

Previous results of current research have been presented in \[9\], \[10\], \[11\]. The key element of our approach is the use of reference samples of real instruments, called instrument prints. The spectral energy of the original recording is split between the notes to be separated based on these prints.

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 shows an overview of the separation process. Descriptions of the building blocks are given in order to let the reader get familiar with the approach. Figure 1 shows the block-diagram of the sound separation process.

The notation of Figure 1 uses the following signs for different representations of signals:

- Simple waveform (W): time domain function of the signal
Instruments of the recording

<table>
<thead>
<tr>
<th>W</th>
<th>FFT</th>
<th>Freq Est</th>
<th>S</th>
<th>Bandogram c.</th>
<th>Sample store</th>
<th>Playmode and volume detector</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>+</td>
<td>F</td>
<td>F</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

User input

<table>
<thead>
<tr>
<th>W</th>
<th>FFT</th>
<th>Freq Est</th>
<th>S</th>
<th>Bandogram c.</th>
<th>Sample store</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td></td>
<td></td>
<td>F</td>
<td>B</td>
<td></td>
</tr>
</tbody>
</table>

Instrument notes for sampling

<table>
<thead>
<tr>
<th>W</th>
<th>FFT</th>
<th>Freq Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 [7], [8] covered different transformation methods in order to determine the best possible means for analysis of audio signals. Current research has examined the analysis of polyphonic musical signals in particular [9], [10].

In [12] 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 $\phi_{k,t}$ respectively. In this document the time index will be omitted in some of the equations for understandability. 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 $f_{true}^{true}$ can be computed for each bin as follows. Let the frequency of the $k$th bin be

$$f_k = k \frac{\text{samplerate}}{\text{framesize}}.$$  

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

$$f_{k,t}^{true} = f_k + \frac{\phi_{k,t}^{true}}{2\pi} \cdot (t_2 - t_1),$$

where

$$\phi_{k,t}^{true} = \phi_{k,t} + (t_2 - t_1) \cdot 2\pi f_k,$$

$$\phi_{k,t}^{dev} = \phi_{k,t} - \phi_{k,t}^{true} + 1 \cdot 2\pi,$$

where $\phi_{k,t}$ is the phase of bin $k$ in time $t$; $\phi_{k,t}^{true}$ is the expected phase; $\phi_{k,t}^{dev}$ is the deviance between the expected and measured phase; $f_{k,t}^{true}$ is the estimated true frequency of bin $k$ in time $t$ and $l \in \mathbb{Z} : -\pi < \phi_{k,t}^{dev} \leq +\pi$.

It is important to mention that, although this method is a very effective tool of signal analysis, it is not used in the transformation back to time domain.

The following subsections describe one block each in detail from Figure 1.
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 detail in Section 3.

2.3 Playmode and volume estimator.
While the user can input the location of the instrument notes in frequency and time, he/she 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 will also be 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, an \( M \) matrix 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 heart 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, although lowers the quality, makes it possible to carry out the separation even under these circumstances. 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 \( t = r \tau \), where \( r \) stands for the current frame and \( \tau \) is the time difference between subsequent frames the original separation problem can be drafted as

\[
\mathbf{c}_r = \sum_{s_i} \mathbf{s}_{s_i}^{orig} \tau,
\]

where \( \mathbf{c}_r = [c_{i,1}, c_{i,2}, \ldots, c_{i,J}] \) is the input signal which is the input of the separation algorithm and \( \mathbf{s}_{s_i}^{orig} = [s_{i,1}, s_{i,2}, \ldots, s_{i,J}] \) are the original notes. This undetermined system of equations cannot be solved unambiguously without other constraints to add.

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 me made, among which the most obvious change is to eliminate the unknown \( \hat{\sigma}_{i,k,\tau} \) phases from the equation system:

\[
\hat{\sigma}_{i,k,\tau} = \gamma_{i,k,\tau}.
\]

This rephrases the original problem to

\[
\mathbf{c}_r = \sum_{s_i} \hat{s}_{s_i,\tau} + \hat{\mathbf{c}}_r,
\]

where \( \hat{s}_{s_i} \) stands for separated note \( s_i \) and \( \hat{\mathbf{c}}_r \) is the remaining energy in the recording after the separation. This perceptually motivated modification exploits the fact that the human ear does 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 showed 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 that 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 as 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 when 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 \( f_{\text{base}} \), called the fundamental frequency. Their time function can be expressed as

\[
x(t) = \sum_{p=1}^{P} a_p \cdot \sin(2\pi \cdot p \cdot f_{\text{base}} \cdot t + \phi_0).
\]

where \( a_p \) is the amplitude of the \( p^{\text{th}} \) partial, \( P \) is the total number of partials and \( \phi_0 \) 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 missing fundamental) [13] (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

\[
\frac{f_2}{f_1} = 2^{\frac{1}{12}}.
\]

An interval of \( n \) semitones is defined by

\[
\frac{f_2}{f_1} = 2^{\frac{n}{12}},
\]

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

\[
 n = 12 \cdot \log_2 \left( \frac{f_2}{f_1} \right)
\]

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

<table>
<thead>
<tr>
<th>Harmonic</th>
<th>Freq</th>
<th>Approximate interval with ( f_{\text{base}} )</th>
<th>Pitch class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( f_{\text{base}} )</td>
<td>unison</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>( 2f_{\text{base}} )</td>
<td>octave</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>( 3f_{\text{base}} )</td>
<td>octave + 5th</td>
<td>E</td>
</tr>
<tr>
<td>4</td>
<td>( 4f_{\text{base}} )</td>
<td>2 octaves</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>( 5f_{\text{base}} )</td>
<td>2 octaves + major 3rd</td>
<td>C#</td>
</tr>
<tr>
<td>6</td>
<td>( 6f_{\text{base}} )</td>
<td>2 octaves + 5th</td>
<td>E</td>
</tr>
<tr>
<td>7</td>
<td>( 7f_{\text{base}} )</td>
<td>2 octaves + 7th</td>
<td>G</td>
</tr>
<tr>
<td>8</td>
<td>( 8f_{\text{base}} )</td>
<td>3 octaves</td>
<td>A</td>
</tr>
</tbody>
</table>

Table 1: Intervals between the first 8 harmonics of a complex tone and its fundamental frequency \( f_{\text{base}} \).

Example for the harmonics of A

![Figure 2: Harmonic series from A2](image)

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 will always be tuned to this pitch.

The importance of understanding the described structure of instrument notes clearly shows 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 is very likely to coincide with the base tone or a harmonic of another one. In these cases the energy splitter needs a hint on how big amount 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 input 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. However, as said earlier, we may not be able to separate e.g. two notes on the same frequency without any information of their properties. For this reason we will store the representation of instruments. This representation will be called an instrument print. [14] 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 make a human listener able to distinguish between them. The instrument prints in this paper are in part based on these researches, in the sense that they contain the features that were found important in the experiments mentioned. 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 an 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 $M=[m_1,m_2,...,m_r]$. It can be regarded as a function

$$A(M,f,b,\tau,t)$$

with the conditions

$$t,m,b,\tau = R^+$$

$$0 < m < m_{max}$$

$$0 \leq t < \infty$$

$$0 < f \leq 20000 Hz$$

that shows how amplitudes through the frequency range change over time for a specific note at a specific frequency $f_{base}$ played with a specific playmode $M$.

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’. 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

$$A_{M,f_{base},k,\tau} = \sum_{f_{base} < f < f_{max}} c_{k,\tau}$$

where

$$b = \log\frac{f_{base}}{f_{max}}$$

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.

As said 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
4 Conclusion

The paper has shown a method for separating instrument notes in recording using pre-recorded instrument prints. The results are quite promising. Examples of separation 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 on each other’s base or overtone frequencies the separation provides reasonably good results.

Acknowledgements

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

References


