SEPARATION OF HARMONIC INSTRUMENTS WITH OVERLAPPING PARTIALS IN MULTI-CHANNEL MIXTURES

Harald Viste

Audiovisual Communications Lab
Swiss Federal Institute of Technology Lausanne
Switzerland
harald.viste@epfl.ch

Gianpaolo Evangelista

Dep. of Physical Sciences
University of Naples "Federico II"
Italy
gianpaolo.evangelista@na.infn.it

ABSTRACT

When instruments play together, different partials will often overlap in time and frequency. This is particularly likely for harmonic instruments. We present a new method for the separation of overlapping partials in multi-channel mixtures. The method is based on the observation that when a harmonic instrument plays a note, all the partials have similar shapes, i.e. common onset, offset, amplitude and frequency modulation. For narrow band partials we devise a method to estimate a demixing matrix that can recover the original source partials from the multi-channel mixture where they overlap. The method is computationally efficient in that it works on highly downsampled narrow frequency bands and it performs equally well for closely spaced partials as for crossing partials, e.g. due to frequency modulations such as vibrato effects. It is able to separate partials in mixtures with a high number of overlapping partials, such as two instruments playing notes where the fundamental frequencies are in fifth (3:2) or octave (2:1) relation.

1. INTRODUCTION

The problem of separating individual sound sources from a mixture of these is one of the core problems of Computational Auditory Scene Analysis (CASA). This problem has become increasingly popular over the recent decades and a number of methods have emerged. However, none of these deals successfully with mixtures of harmonic instruments where many of the partials overlap.

In sinusoidal models [1] each sound is represented as a set of sinusoidal trajectories. Each of these is parametrized by its amplitude, frequency and phase trajectories. Sinusoidal modeling is a well suited tool for the detection of overlapping partials, but it is only able to separate partials that can be detected as separate trajectories in the time-frequency representation. There exist several methods to improve the frequency resolution of the FFT [2]. A model fitting approach for least-squares estimation of colliding sinusoids [3] and methods for interpolation of colliding trajectories [4] have also been proposed. None of these methods come close to the frequency resolution that may be needed in order to separate overlapping partials from harmonic instruments. Sinusoidal models have also been used for cancellation of beatings in closely spaced partials [5]. However, all of these approaches have problems in capturing small frequency fluctuations in the original partials. This means that, e.g., vibrato is not preserved in the separated signals and small errors in the partial frequency may bring them "out of tune".

Multi-channel blind source separation (BSS) [6, 7] is another group of techniques that can be used for partial separation. These are iterative methods that work under the assumption that all the sources are statistically independent. It may be questioned whether this assumption holds for overlapping partials in mixtures of harmonic instruments. Though these methods may work well for broad band signals, our experiments with such methods on narrow band overlapping partials in isolation have shown convergence problems, and accordingly no proper separation.

We devise a new method for energy separation of overlapping partials in multi-channel mixtures, drawing benefits from sinusoidal modeling and multi-channel techniques. This method works individually on partials, it is computationally fast, and it can be used in conjunction with many existing source separation methods.

In this paper, first we describe our method in section 2. Then, section 3 discusses some experimental results, followed by section 4 on measurements of the separation quality. In section 5 we draw the conclusions.

2. SEPARATION OF OVERLAPPING PARTIALS

2.1. Signal representation

In a general source separation setup, the only known signals are the sensor signals $x_m[t]$, where m is the sensor index. We use the short-time Fourier transform (STFT) in order to represent the signals in time and frequency $X_m[\omega, t]$. Each of these signals is a superposition of filtered source signals. In a general scene with M sensors and N sources, this can be written in matrix notation as $\mathbf{X} = \mathbf{HS}$:

$$\begin{bmatrix} X_1[\omega,t] \\ \vdots \\ X_M[\omega,t] \end{bmatrix} = \begin{bmatrix} H_{11}[\omega] & \cdots & H_{1N}[\omega] \\ \vdots & \ddots & \vdots \\ H_{M1}[\omega] & \cdots & H_{MN}[\omega] \end{bmatrix} \begin{bmatrix} S_1[\omega,t] \\ \vdots \\ S_N[\omega,t] \end{bmatrix} \quad (1)$$

where $S_n[\omega, t]$ are the source signals and H is the $M \times N$ mixing matrix. Each of the elements $H_{mn}[\omega]$ of this matrix is a filter that describes the filtering between the n'th source and the m'th sensor. We assume that the scene is static and that these filters are time invariant.

The STFT representation is well suited for the estimation of parameters in sinusoidal models and we use this property in order to detect partials in the signals. However, a sinusoidal trajectory is simply a single sinusoid with time-varying frequency (and amplitude), corresponding to a thin line in the time-frequency representation. This only describes the ridge of a partial and contains no

0-7803-7850-4/03/$17.00 © 2003 IEEE

information about the sidebands, i.e. the slopes at the frequencies below and above. These sidebands are important for the naturalness of a sound. Moreover, in a frequency band that contains several overlapping partials, there will be several such ridges that may be impossible to distinguish from each other. We therefore need to extract and model signal components that contain both the partial ridges and their sidebands. In contrast to traditional sinusoidal modeling, we are not estimating a very accurate frequency trajectory for the ridge, but rather a rough region in the time-frequency representation that will contain the ridge and its sidebands.

Effectively we divide the entire time-frequency plane into non-overlapping regions Ω_i indexed by i. Each such region can be described by the corresponding indicator function $I_i[\omega,t]$ which equals 1 for $[\omega,t] \in \Omega_i$ and 0 elsewhere. Due to the typical shapes of these regions and the fact that each such region may contain a mix of many partials from different sources we will refer to these as "partial sausages", or simply sausages. Since the sausages are non-overlapping and span the whole time-frequency plane, each of the sensor signals may be written as a sum of **sensor sausages**:

$$X_m[\omega,t] = \sum_i P_{im}[\omega,t] \tag{2}$$

where each sensor sausage is simply the product of the corresponding sensor signal and the indicator function:

$$P_{im}[\omega,t] = I_i[\omega,t]X_m[\omega,t] \tag{3}$$

Basically, each sensor sausage may either contain partial(s) from one source, or a mix of overlapping partials from several sources.

We employ grouping principles based on harmonic relations [8] and localization cues [9] in order to determine the total number of sources N, as well as to detect which sources that contribute energy in each of the sausages. We thus have a mapping between sausages and sources. For each of the sources we then know which sausages contain partials from that particular source. Clearly, if each of the sensor sausages only contains partials from one of the sources, separation is straightforward. The time-frequency energy of each sausage is simply assigned to that corresponding source. Any source is then the union of the sausages that contain its partials.

When the sausages contain overlapping partials from different sources we need to decompose each sensor sausage into its source components, namely the original **source sausages** or source partials that the sensor sausage is a superposition of. Separation is then achieved by working with these source sausages rather than the mixture sausages, i.e. we assign these source sausages to the different sources.

Combining (1) and (3) we get

$$\begin{bmatrix} P_{i1}[\omega,t] \\ \vdots \\ P_{iM}[\omega,t] \end{bmatrix} = I_i[\omega,t] \begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & \ddots & \vdots \\ H_{M1} & \cdots & H_{MN} \end{bmatrix} \begin{bmatrix} S_1[\omega,t] \\ \vdots \\ S_N[\omega,t] \end{bmatrix} \tag{4}$$

For most harmonic instruments the partials are narrow band. This means that over the frequency range of one sausage the mixing filters $H_{mn}(\omega)$ can be assumed as constant complex numbers. We therefore omitted the parameter ω in (4). We denote the constant mixing matrix corresponding to the i'th sausage by $\mathbf{H}_i$.

When the number of sources that are overlapping in a sensor sausage is not greater than the number of sensors and the corresponding mixing vectors (rows of $\mathbf{H}_i$) are linearly independent it is possible to separate the overlapping partials in the sensor sausages to obtain the original source partials. By estimating the left pseudo-inverse $\mathbf{G}_i$ of $\mathbf{H}_i$ and applying this on both sides of (4) we get:

$$\begin{bmatrix} R_{i1}[\omega,t] \\ \vdots \\ R_{iN}[\omega,t] \end{bmatrix} = \mathbf{G}_i \begin{bmatrix} P_{i1}[\omega,t] \\ \vdots \\ P_{iM}[\omega,t] \end{bmatrix} \approx I_i[\omega,t] \begin{bmatrix} S_1[\omega,t] \\ \vdots \\ S_N[\omega,t] \end{bmatrix} \tag{5}$$

where R_{in} are the **separated sausages**. Each R_{in} then represents the contribution of a single source S_n in the sausage region I_i. To find the separated sausages, we need to estimate the matrix $\mathbf{G}_i$ for each of these sausage regions.

2.2. Similarity of partial envelope shapes

From psychophysics it is known that the human hearing sense uses many cues in order to group together different simple sound components into one complex sound. In particular, looking at the partials of one single note from a harmonic instrument, one can consider the harmonic relation, the common onset, offset, amplitude modulation (AM), and frequency modulation (FM). These are all important cues for grouping.

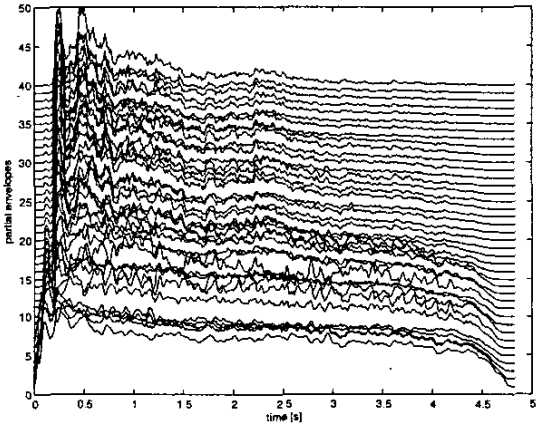

Figure 1: *Energy envelopes of first 40 partials of a trumpet note.*

For any sausage, let us denote it A, we define the energy envelope $E_A[t]$ and normalized energy envelope $\bar{E}_A[t]$ as follows:

$$E_A[t] = \sum_\omega A[\omega,t]^2 \quad , \quad \bar{E}_A[t] = \frac{E_A[t]}{\|E_A[t]\|_2} \tag{6}$$

The sausage energy envelope (6) preserves all of the abovementioned cues except for the FM. Figure 1 shows the normalized energy envelopes of the first 40 partial sausages of a trumpet note. The graphs are slightly shifted vertically in order to better show the resemblances.

From the mapping we have found between the sensors and the sausages we know which sausages that contain partials for any given source. Since all the partial envelopes of a harmonic sound have similar shapes, any sensor sausage containing energy from only one source can be used in order to predict the envelopes for all the partials of that corresponding source. We call these the **predicted sausages** Q_{in}, The envelopes $\bar{E}_{Q_{in}}$ of these are used

as predictions for the envelopes $\bar{E}_{R_{in}}$ of the separated sausages R_{in} that we are looking for. Using (5) we find the $\mathbf{G}_i$ that gives R_{in} with envelopes $\bar{E}_{R_{in}}$ as close as possible to the predicted envelopes $\bar{E}_{Q_{in}}$'s. We define the envelope shape similarity β be-

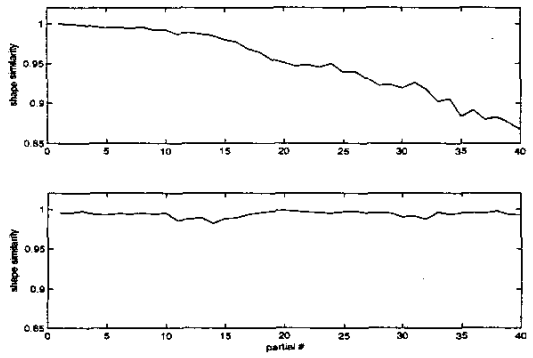

Figure 2: *Partial envelope similarity relative to the first partial (top), and relative to the fifth neighbour partial (bottom).*

tween two sausages A and B as the inner product of their normalized energy envelopes:

$$\beta(A,B) = \langle \bar{E}_A[t], \bar{E}_B[t] \rangle = \sum_t \bar{E}_A[t]\bar{E}_B[t] \qquad (7)$$

This measurement is in the range between 0 and 1, where 1 means that the two normalized envelopes are identical. Figure 2 shows this similarity measure for the 40 first partials shown in figure 1. The top graph shows the similarity of each partial relative to the first partial, $\beta(S_i, S_1)$. The bottom plot shows the similarities between the 40 partials and their 5th neighbour partials, $\beta(S_i, S_{i+5})$. Even though the similarity relative to the first partial decreases as the partial index increases, the local similarity remains high. Of course this depends on the type of instrument and note being played, but for harmonic instruments it is reasonable to assume a significant correlation between the envelopes of the different partials of a note.

For a given sausage index i we find the $\mathbf{G}_i$ that maximizes the total sausage shape similarity measure given by:

$$\beta_i = \sum_n \beta(Q_{in}, R_{in}) \qquad (8)$$

Using this $\mathbf{G}_i$ in (5) the sensor sausages P_{im} are separated into source sausages R_{in}.

If the sensors are placed in a free field with little reverberation, then the mixing filters will be approximately pure delays. In this case it is possible to estimate the complex elements of the matrix $\mathbf{H}_i$ for each sausage, drawing knowledge from the prediction sausages. In situations with no reverberation at all, one can directly compute the (pseudo-)inverse $\mathbf{G}_i$ and use this directly in (5). In more complex scenes, we use this rough estimate of $\mathbf{H}_i$ as the starting point for the iterative algorithm that maximizes the similarity (8).

3. RESULTS

Figure 3 shows the (unnormalized) energy envelopes for a sausage region containing two overlapping partials, coming from a violin with vibrato and a trumpet. The top graph shows the envelopes E_P of the two sensor sausages (left and right sensor). The second row shows the envelopes E_Q of the two predicted sausages. These envelopes are deduced from neighbouring non-overlapping sensor sausages. The third row shows the envelopes E_R of the separated sausages. Finally, the bottom graphs show the envelopes E_S of the original source sausages $I_i[\omega,t]S_n[\omega,t]$, which represent the perfect separation. Of course these two latter are only known when one knows the original source signals, as in our research setup. They are shown here just for comparison. Surprisingly, we note that the separated sausages (third row) are better (both scale and shape) than we could expect. In other words, they are closer to the perfect source partial envelopes (bottom) than to the predicted envelopes we were looking for (second row). In particular we notice that the vibrato and strong amplitude modulation of the violin has been preserved (left), whereas this has been correctly removed for the trumpet (right).

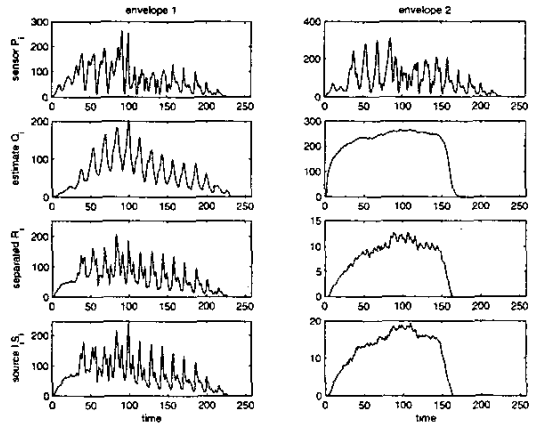

Figure 3: *Sausage envelopes, from top to bottom: sensor sausages containing overlapping partials (P's), predicted sausages from neighbouring non-overlapping partials (Q's), separated sausages (R's), and original source sausages (S's).*

So far we have only been concerned with the energy envelopes of the sausages. Since we are working in the STFT domain all the sausages are complex. What is actually much more important for the perceptual quality of the separation is the phase trajectory of the separated partials. Partials that are slightly out of tune are typically much more annoying than fluctuations in the energy envelopes. Figure 4 shows phase differences relative to the original source sausages that represent perfect separation. The figure layout is the same as in fig. 3. In the top row we see that the sensor signals have quite random phase in relation to the original sources. This is expected since there are several overlapping partials. The second row shows the phase difference between the predicted sausages and the source sausages. Naturally, these predicted sausages have been deduced from sausages that lie in different frequency bands. At best the phase difference is linear (taking into account the phase wrapping at $\pm\pi$) as seen in the right hand graph. This depends on the individual instrument and mixing, and is therefore not useful in general. The third row shows the phase difference between the separated sausages and the original source sausages. We see that the phase is almost constant. This means that

there are almost no phase distortions (except for a pure delay). The large phase error seen at the end of the right figure corresponds to a time interval where the source is silent. This means that the phase error is less significant in this interval. At the bottom we show the phase errors of the perfect source sausages relative to themselves (zero by definition).

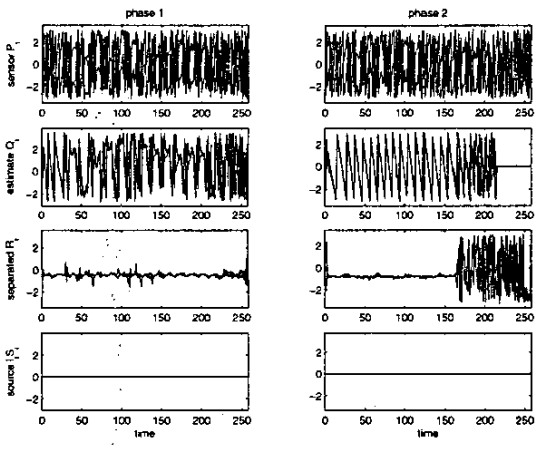

Figure 4: *Phase errors from top to bottom: sensor sausages (P's), predicted sausages (Q's), separated sausages (R's), and original source sausages (S's).*

4. QUALITY MEASURES

It turns out that for the perceived quality of separation, the accuracy of the sausage envelopes is not the most important. As a matter of fact, a partial that is much too strong (scaling error) or slightly out of tune (phase/frequency error) is normally perceived as much more annoying than changes in the shape of its normalized energy envelope. However, it is possible to predict the normalized envelope, and this is why we use this in the similarity measurement (8) in order to demix the overlapping partials. The hope is that the scaling and phase of the Q_{in} that we get from (5) are correct to a good degree of approximation, as seen in fig. 3 and 4 respectively.

4.1. Scaling error

Depending on the physics of a harmonic instrument, the partials have different energy levels. However, these levels are normally somewhat interrelated, and normally follow a general trend. The level typically decreases with increasing frequency, and one partial is not likely to be very much stronger than the neighbouring partials.

With a properly chosen windowing function in the STFT, the total energy of a sausage A is given by

$$\epsilon_A = \sum_t \sum_\omega |A[\omega, t]|^2 \tag{9}$$

By comparing the strength of all the partials in the separated signals we detect separated partials that are clearly wrong and use another method such as partial shape smoothing or partial cancellation [5] for the sausage under consideration.

4.2. Phase error

We denote the phase difference between two sausages A and B as $\phi_{AB}[\omega, t]$. The energy weighted mean of this error is:

$$\bar{\phi}_{AB} = \frac{1}{\epsilon_A} \sum_\omega \sum_t \phi_{AB}[\omega, t] |A[\omega, t]|^2 \tag{10}$$

In research experiments where the original source signals are available, the variance of the phase error ϕ_{RS} between a separated sausage R and a corresponding source sausage S can be used as a quantitative measurement for the quality of this separation. In our experiments we have achieved phase errors with very small variance as seen in fig. 4. Partials containing vibrato can effectively be separated from partials without frequency modulations.

5. CONCLUSIONS

We have presented a new method for separation of overlapping partials in multi-channel audio mixtures. This method can accurately recover the amplitude and frequency modulation of the original sources from the mixtures. It can be used in conjunction with existing source separation methods. Sound examples can be found at: `http://lcavwww.epfl.ch/~viste/waspaa03`

6. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech analysis/sysnthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[2] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proceedings of 5th International Conference on Digital Audio Effects*, 2002, pp. 51–58.

[3] T. Tolonen, "Methods for separation of harmonic sound sources using sinusoidal modeling," in *the AES 106th Convention, Munich, Germany*, 1999.

[4] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 2000, pp. 765–768.

[5] H. Viste and G. Evangelista, "An extension for source separation techniques avoiding beats," in *Proceedings of 5th International Conference on Digital Audio Effects*, 2002, pp. 71–76.

[6] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Int. Workshop on Independence and Artificial Neural Networks*, 1998, pp. 19–25.

[7] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proceedings of International workshop on Independent Component Analysis and Blind Signal Separation*, 1999, pp. 365–371.

[8] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 2002, pp. 1757–1760.

[9] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 2000, pp. 2985–2988.