A Method for Separation of Overlapping Partials Based on Similarity of Temporal Envelopes in Multi-Channel Mixtures

Harald Viste* and Gianpaolo Evangelista

Submitted to: **IEEE Transactions on Speech and Audio Processing**
EDICS Category: 2-MUSI Signal Processing for Music

Date: February 17, 2005

***Abstract*— In a situation where multiple sound sources are concurrently active, the signals of the individual sources often overlap in time and in frequency. This is particularly likely for voiced instruments where the frequencies of some of the partials of one single note coincide with the frequencies of some of the partials of another instrument playing a harmonically related note. A source separation algorithm suitable for musical applications must address the problem of overlapping partials.**

A method is proposed for the separation of overlapping narrow-band partials in multi-channel mixtures. The method is based on the observation that, for many instruments, all the partials of a single note have similar temporal envelopes. For narrow band partials these similarities can be exploited in order to estimate demixing matrices in the frequency domain. Effectively, one can recover estimates of the original partials from a multi-channel mixture where they overlap. The method is computationally efficient in that it works on highly downsampled narrow frequency bands. It performs well for closely spaced and colliding partials, and (to some extent) also for frequency modulations such as vibrato effects.

***Index Terms*— source separation, harmonic instruments, overlapping partials.**

I. Introduction

MANY INSTRUMENTS exhibit a strong sinusoidal nature. For a single excitation, or single note, the corresponding signal can be closely modeled as a sum of sinusoids whose amplitudes and phases vary slowly in time. In a suitable time-frequency representation these sinusoids can be detected, and the parameters of the amplitude and phase trajectories can be estimated as functions of time [1]. These sinusoidal components are, in general, called *partials*, since each of them constitutes an important part of the total signal. When instruments play together in an ensemble there is a high chance that their energy distributions overlap in time and in frequency. Depending on the types of instruments being played, and on the tuning of the various instruments, some of the partials of one instrument may overlap with the partials of other instruments.

Several techniques for the separation of audio sources have been proposed and studied in the literature. Many of these work well for certain types of audio signals, such as speech, whereas only few of them deal specifically with harmonic instrument and music signals. In the following some of the existing techniques are reviewed.

A. Separation by source properties

Some of the earliest techniques for source separation are based on sinusoidal modeling. In sinusoidal models [1], each partial is described as a sinusoidal trajectory, parametrized by amplitude, frequency, and phase as functions of time. These techniques have shown to be powerful for low bit-rate coding of both speech and instruments. Only one sensor signal is typically needed, and time-frequency analysis and spectral peak picking techniques are employed in order to estimate the individual partials. Separation is then achieved by applying grouping principles on the identified partials [2], [3]. Time-frequency analysis is a well suited tool for the detection of overlapping partials. However, due to the time-frequency uncertainty it may not be able to detect the individual partials that overlap as separate trajectories.

Some attempts have been made to resolve closely spaced partials. These include curve fitting techniques and least-squares estimation of closely spaced sinusoids [4]–[7], interpolation and approximation of colliding trajectories [2], [5], [8], and frequency resolution enhancement techniques [9]. However, none of these techniques are able to correctly resolve the amplitude and phase trajectories of the individual partials. In fact, frequency modulation, e.g. vibrato, is not preserved. Even small errors in the estimated frequency of a partial may cause that partial to sound harsh, or “out of tune”, with respect to the other partials of the sound.

B. Separation by mixing model properties

More recently, with the availability of more processing power, multi-channel source separation techniques have become popular. Typically, these techniques do not depend on any specific model for the source signals, but rather exploit

H. Viste is with Audiovisual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, CH-1015 Lausanne, Switzerland (e-mail: viste@lcavsun1.epfl.ch). G. Evangelista is with Dept. of Physical Sciences, University of Naples “Federico II”, via Cintia, 80126 Napoli, Italy (e-mail: gianpaolo.evangelista@na.infn.it).
This work was supported (in part) by the Swiss National Science Foundation under grant number 21-57220.99.

properties of the mixing process. In other words, spatial properties inherent to the sensor signals are used, implicitly or explicitly, in order to separate sources at different physical locations.

Beamforming techniques [10], [11] make explicit use of spatial cues in order to design or estimate spatial filters, i.e. filters that retain signals coming from a particular direction while suppressing signals from all other directions. Time-domain beamforming can achieve high spatial resolution, but only in quite narrow frequency bands. For broadband signals, the problem can be transformed into the frequency domain, where each frequency band can be processed independently. This enables a high spatial resolution for broadband signals as well. However, spatial filtering and temporal filtering are not independent. This means that when two sources at different locations have overlapping partials, the length of the filters needed for proper separation may exceed the length of the signal being analyzed. This is analogous to the problem seen in methods based on sinusoidal models.

Another type of technique is based on time-frequency weighting [12]–[14], which can be seen as a generalization of frequency domain beamforming. The sensor signals are analyzed in time and frequency, typically using a Short-Time Fourier Transform (STFT). Spatial cues, such as level and phase differences, are estimated in this representation. The sources can then be localized by detecting clusters in the spatial cue parameter space. Separation is effectively achieved by generating a time-frequency weighting mask for each source (cluster), and multiplying the STFT of the sensor signals with this mask. In other words, the energy of each time-frequency bin is distributed among all the sources. This distribution is determined by the weights, which are computed based on the distances between the spatial cues of these bins and the spatial cues of the cluster centers.

Blind source separation methods are iterative methods working under the assumption that all the sources are statistically independent. These methods iteratively search for the demixing matrix that best decorrelates the sources. The spatial information in the mixing process is not used explicitly. Techniques that work in the time domain [15]–[18] need to estimate quite long mixing filters, and are computationally expensive. Since these techniques work on the entire signals, they are able to separate overlapping partials, implicitly. However, they are not applicable to individual frequency bands, and generally there cannot be more sources than sensors. It is possible to transform the problem into the frequency domain [19]. This decreases the computational complexity at the cost of introducing permutation ambiguities. Some approaches for solving these exist [20]. In order to have sufficient data for the statistical analysis, each of the frequency bands needs to retain a high number of data points (samples). Consequently, the individual frequency bands cannot be downsampled by a large factor, so that the computational cost remains relatively high. When partials overlap in narrow frequency bands, the assumption about statistical independence may not hold in each individual frequency band. For stationary sounds with smooth partial envelopes, overlapping partials are correlated. In such cases the statistical methods may not converge.

C. Proposed separation technique

We propose a method for separation of overlapping partials in multi-channel mixtures. The method consists of an analysis stage and a separation stage. In the analysis stage, distinct regions in the time-frequency plane are identified, denoted "partial regions", such that each of these regions covers one or more partials. Furthermore a mapping between these and the various sources is established. This is achieved by employing grouping principles on both spatial cues (multi-channel) and partial cues (single channel). The information provided by the analysis is then exploited in the separation stage. Each partial region contains either overlapping partials or single (non-overlapping) partials in frequency regions where the source do not overlap. The partial regions that contain single partials provide information about the characteristics of the underlying sources. The separation uses this information and applies an iterative search for a frequency domain demixing matrix. When applied on regions that contains overlapping partials, this matrix yields separated partials whose envelopes resemble the envelopes of the given non-overlapping partials.

The structure of the paper is as follows. Section II introduces the model for multi-channel source mixing, as well as some considerations about instruments and overlapping partials. This is followed by Section III in which our method is presented. Section IV shows some experimental results. Finally, conclusions are drawn in Section V.

II. Background

A. Multi-channel mixing model

In a general source separation setup the only known variables are the sensor signals $x_m(l)$, where $m = 1..M$ is the sensor index, and l is the time index. In a general setup with M sensors and N sources, each of the sensors records a superposition of filtered source signals. This can be modeled as

$$x_m(l) = \sum_{n=1}^{N} h_{mn}(l) * s_n(l), \tag{1}$$

where '$*$' denotes convolution, x_m denotes the signal at sensor m, s_n are the source signals, and h_{mn} are the filters modeling the sound propagation from source n to sensor m, including the direct path and room reflections.

By employing the Short-Time Fourier Transform (STFT), the signals are represented in time and frequency, and the mixing model can be approximated by:

$$\begin{bmatrix} X_1(k,q) \\ \vdots \\ X_M(k,q) \end{bmatrix} \approx \begin{bmatrix} H_{11}(q) & \cdots & H_{1N}(q) \\ \vdots & \ddots & \vdots \\ H_{M1}(q) & \cdots & H_{MN}(q) \end{bmatrix} \begin{bmatrix} S_1(k,q) \\ \vdots \\ S_N(k,q) \end{bmatrix}, \tag{2}$$

where $X_m(k,q)$ and $S_n(k,q)$ are the STFT spectra of the source and sensor signals, respectively, and $H_{mn}(q)$ are the discrete Fourier transforms of the mixing filters $h_{mn}(l)$, which are assumed to be time-invariant. The arguments k and q are time frame and frequency indexes, respectively. This mixing model can be written in matrix notation as $\mathbf{X}(k,q) = \mathbf{H}(q)\mathbf{S}(k,q)$. In this general framework, source separation is

equivalent to the problem of estimating the mixing matrix $\mathbf{H}(q)$ (or its inverse) for each frequency q. This is explained in more details in Section III.

B. Overlapping energy in music signals

It is important to note that the notion of "overlap" depends on the type of time-frequency analysis being applied. For instance for long stationary sounds, where some partials are closely spaced, it may be possible to sacrifice time resolution in favor of frequency resolution in order to detect the individual partials. Obviously, this is not practical for partials that are only a few Hz apart, since the typical duration of a single note is shorter than the analysis time interval length needed for sufficient frequency resolution. Even worse, when there are frequency fluctuations, such as vibrato, the partial trajectories of different notes and instruments may cross each other. Due to the uncertainty principle, no time-frequency analysis will be fine enough to detect those individual partials accurately.

Harmonic instruments are instruments where the frequencies of the different partials of a single note are in a harmonic or quasi-harmonic relation to each other. In other words, the frequency of each partial is approximately an integer multiple of the fundamental frequency of the corresponding note. In this case the partials are typically called *harmonics*. Other instruments such as drums, bells, but also the piano in the low register, feature partials that are not harmonically related. Throughout this paper the general term partial is used in order to keep the matter as general as possible.

Two tones having a frequency relation that is a ratio of small integers are known to produce a pleasing harmony. A study of psychoacoustical explanations for this "pleasingness" can be found in [21]. In many music genres the most commonly used intervals are those with high consonance. The frequency relations for these intervals are approximately ratios of small integers, such as a fifth (3:2), a third (5:4), a forth (4:3), and an octave (2:1). Consequently, when harmonic instruments play together, it is very likely that some of their partials will overlap.

III. Partial separation

A. Motivation

In the human auditory system, aspects of both groups of methods discussed in Sections I-A and I-B are exploited in order to analyze an auditory scene. Since a person has two ears he/she is able to focus on sounds coming from particular directions. In addition, as a complement, the person is also able to exploit the structure of the individual sources in time and in frequency. It therefore seems plausible to combine aspects of these methods, i.e. spatial analysis and time-frequency analysis.

Partial fusion denotes the phenomena when several partials are perceived as one single auditory event, or as a single note. Some of the cues that are important in order to achieve partial fusion are harmonicity and synchronicity of onset, offset, frequency and amplitude modulation [22, ch.3]. In other words, the different partials of one single note typically have these cues in common. Otherwise, they would not be perceived as one sound. If a few partials for a given note are known, they give a rough idea of what the other partials should resemble. Thus, if some non-overlapping partials can be detected, these can be used as models for the unknown partials that need be separated out of a superposition of overlapping partials. This is the motivation behind the method proposed in this paper. In particular, it can be observed that, for a single note, the temporal envelopes of the partials have quite similar onset/offset times and amplitude modulation. In synthesis techniques this is well known in the form of a model for attack, decay, sustain and release (ADSR model). Column A in Fig. 1 shows the envelopes of the first six partials of a single note being played by an alto trombone. The partial envelopes change somewhat with frequency. For instance, the higher partials have delayed onset times as well as faster decay. However, there is a noticeable similarity between all the partials, and in general the similarity is higher between partials whose frequencies are close to each other.

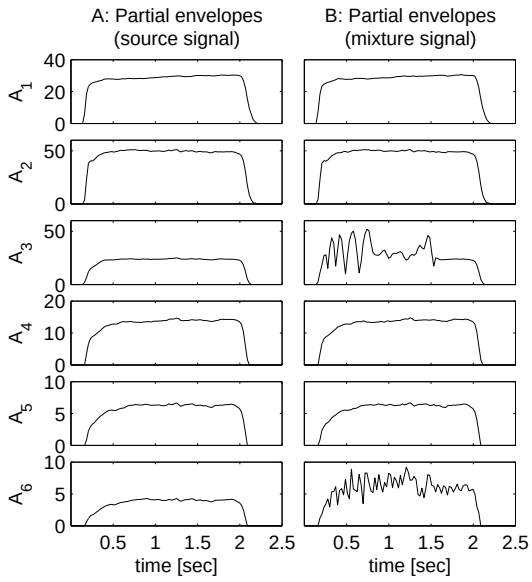

Fig. 1. *Amplitude envelopes of the first 6 partials of an A note played by an alto trombone. A: Original source partials. B: Partials in a mixture, where the third and sixth partials overlap with partials of other instruments.*

The proposed method consists first of a time-frequency analysis where distinct regions, denoted partial regions, are detected in the time-frequency plane such that each region covers the main energy of one or more (overlapping) partials. Then grouping principles based on harmonicity, spatial cues, and temporal envelope shapes of the partials are employed in order to find a mapping between the sources and these regions. This provides information about which sources contribute with significant energy in each of the different regions, as well as which regions contain only one partial (non-overlapping). These latter partials are used as models for the unknown overlapping partials. Finally, for each region containing overlapping partials the demixing matrix is estimated. This is accomplished by an iterative search for the matrix which gives separated partials whose temporal envelope shapes most closely resemble those of the model partials.

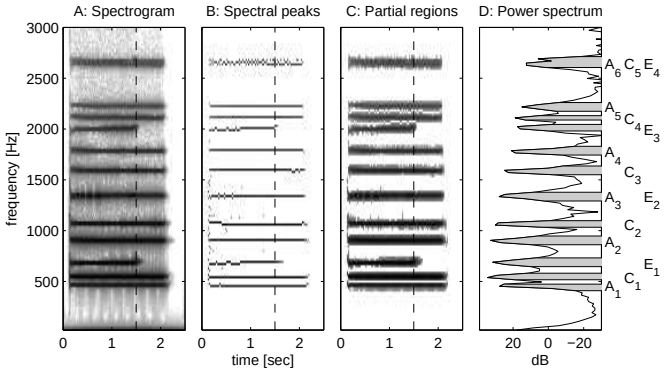

Fig. 2. *Three notes (A minor chord) played by alto trombones, from left to right: A: spectrogram, B: spectral peaks, C: partial regions around the spectral peaks, D: power spectrum and detected partial regions for one particular time frame index (indicated by dashed lines in the other panels).*

B. Definition of partial regions

The STFT is a well suited tool for the detection of partials and for the estimation of the parameters in sinusoidal models. Like in most sinusoidal modeling techniques, the peaks in the STFT are used in order to describe the partial trajectories. However, unlike sinusoidal modeling, we are not trying to characterize each of these partials by their amplitude and phase trajectories. Trying to do this to a high degree of accuracy would be useless since each spectral peak may actually be the net result of several overlapping partials in its close neighborhood. In addition, the sidebands of the partials also contain information that may be important for the naturalness of the sounds. Therefore, the goal is not to estimate the amplitude and phase trajectories to a high degree of accuracy. Rather, the aim is to define regions around the partial trajectories (partial regions), such that any underlying partial is covered by these regions. At each time frame k, for each detected spectral peak, a frequency range is formed by including neighboring frequency bins at lower and higher frequencies, as long as the amplitude is strictly decreasing (up to some maximum region width, e.g. 150 Hz was used in the examples in this paper). These frequency ranges are then connected across time by means of frame-to-frame frequency range matching, yielding partial regions. This is entirely similar to the traditional frame-to-frame peak matching used in sinusoidal models [1], but with the advantage that, for a given time frame k, multiple spectral peaks may be assigned to the same partial region. This is illustrated further in the following example.

Figure 2 shows an example for a signal consisting of three notes, namely the notes A, C, and E, which together constitute an A minor chord. Panel A shows the spectrogram of the signal (one of the sensor channels). The sinusoidal nature of the partials, as well as the harmonic structure, can be clearly seen. Panel B shows the spectral peaks that were detected by peak picking. For the examples in this paper the spectrum of a single sensor channel (the first sensor) was used, and a hard threshold was used for the detection of the spectral peaks in this spectrum. This simple approach worked well in our examples. In practice, however, a more robust peak picking method should be chosen. This can be achieved by taking into account all sensor channels and by employing more sophisticated multipitch estimation techniques [23], [24]. Panel C shows the partial regions that have been identified around these spectral peaks. Panel D shows the power spectrum of the signal for one particular time frame index. Its temporal position is indicated by vertical dashed lines in the three other panels. Around the most prominent peaks in the power spectrum partial regions have been identified. These are shown as horizontal, filled bars. The frequency widths of the partial regions at the given time instant equals the extent of these bars on the frequency axis (ordinate). Finally, the partials are annotated with the names of the underlying notes and their partial indexes, i.e. A_2 denotes the second partial of the A note. The relative fundamental frequencies of the three notes are approximately $\frac{1}{1}$, $\frac{6}{5}$, and $\frac{3}{2}$, respectively. The partials A_3 and E_2 overlap, introducing amplitude fluctuations, denoted beatings. This can be seen as intensity variations at the beginning of the corresponding partial region in panel A in Fig. 2. Similarly, the partials A_6, C_5 and E_4 overlap. The peak picking algorithm is naturally not able to detect each of these individually. This can be seen in panel B, where the multiple peaks do not form one clear partial trajectory. The effect of the overlapping partials can also be observed in column B in Fig. 1 where the envelopes of the partial regions corresponding to the first six partials of the A note are shown. The overlapping partials with strong beatings and erroneous envelopes are clearly seen in the third and sixth panel from the top, respectively. Column A shows the envelopes of the uncorrupted partials of the original source signal in the same partial regions.

In traditional sinusoidal modeling there can be ambiguous situations in which the frame-to-frame peak matching algorithm needs to make a hard decision. When there are several possible candidate peaks for the continuation of a partial, only one of these peaks can be chosen. The other peaks must be either discarded, or modeled as separate partials. The example with the three overlapping partials in Fig. 2 illustrates this. For some time frame k, two or more peaks are detected (panel B) in the narrow band region where these partials overlap (about 2600 Hz). Neither is it possible to connect these peaks into one single partial trajectory (without ambiguities), nor do these peaks provide estimates of the phase and amplitudes of the underlying partials.

When working with regions covering the partial trajectories, rather than trying to accurately characterize them, multiple candidate peaks no longer constitute a problem. Whenever there are two (or more) possible candidate regions for the continuation of a given partial region, any number of these candidate regions, as well as any non-covered bins between them, can be included in the given partial region. In panel D in Fig. 2 it can be seen how two candidate peaks (about 2600 Hz) have been included in the same partial region. In panel C it can be seen how the corresponding partial region effectively covers the multiple candidate peaks.

Partial regions provide a means for dividing the entire time-frequency plane into non-overlapping regions Ω_i, indexed by i, where each region captures the main energy from one or more partials. Each partial region can be described by

the corresponding indicator function $I_i(k,q)$, which equals 1 for $(k,q) \in \Omega_i$ and 0 elsewhere. For a given signal, each such region may contain a single partial or a superposition of overlapping partials. For notational convenience, we use the term partial to denote the energy of any signal that is contained in a single partial region, even when the region actually contains a superposition of two or more overlapping partials. The term *sensor partial* is used to denote the part of a sensor signal that is contained in a given partial region. A sensor partial may consist of one single partial, or a combination of several overlapping partials. Similarly, for the original (unknown) source signals, the term *source partial* is used.

Even if the selected partial regions do not cover the entire time-frequency plane, they capture most of the energy in the signal. It can therefore be assumed that the entire plane is covered such that any residue, or bins that are not part of any partial region, is neglected. The residue is better handled by other separation techniques, such as those based on time-frequency weighting, as discussed in Section I-B. Under the assumption that the entire time-frequency plane is covered, each of the sensor signals may be written as a sum of sensor partials:

$$X_m(k,q) = \sum_i P_{im}(k,q), \tag{3}$$

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

$$P_{im}(k,q) = X_m(k,q) I_i(k,q). \tag{4}$$

C. Partial temporal envelope similarity

The separation method that we propose in this paper is based on the observation that, for many instruments, all the partials of a single note have similar temporal envelopes. In order to determine how similar two partial envelopes are, a measure of partial similarity is needed. For a given partial region Ω_i, the temporal envelopes $E_{P_{im}}(k)$ of the sensor partials $P_{im}(k,q)$ are defined as follows:

$$E_{P_{im}}(k) = \sqrt{\sum_q |P_{im}(k,q)|^2}. \tag{5}$$

The temporal envelope of the signal in a partial region (5) contains information about onset, offset, and amplitude modulation, which are the fusion cues mentioned in Section III-A. It does not contain frequency modulation information.

In order for the temporal envelope to have an intuitive meaning in terms of signal strength (i.e. power) as a function of time, the choice of analysis window $w(l)$ and hop-size L in the STFT is restricted. For any signal $s(l)$ with STFT $S(k,q)$, the sum of local powers must equal the total signal energy (up to some constant scaling factor C):

$$\sum_k \sum_q |S(k,q)|^2 = \sum_k \sum_l [w(l-kL)s(l)]^2 = C \sum_l s(l)^2. \tag{6}$$

This means that the STFT basis functions constitute a tight frame, and yields the following constraint:

$$\sum_k w(l-kL)^2 = C, \quad \forall l. \tag{7}$$

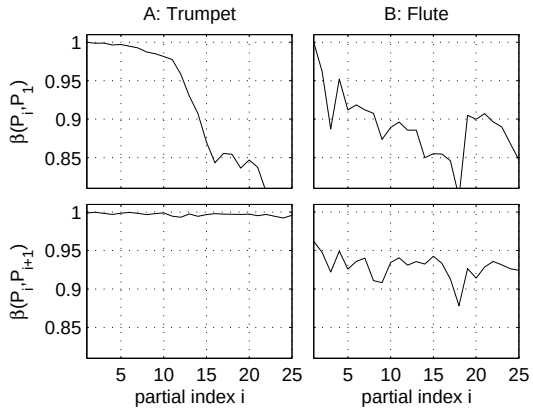

Fig. 3. *Partial envelope similarity for two notes, A: trumpet, B: flute. Top row: similarity relative to the first partial, Bottom row: similarity relative to the neighbor partial.*

For the examples in this paper, we have used the half-wave sine window, with 50% overlap. A window length of about 50 ms was chosen.

The temporal envelope shape similarity β between two partials, $P_{im}(k,q)$ and $P_{jm}(k,q)$, can be defined as the inner product between their normalized temporal envelopes:

$$\beta(P_{im}, P_{jm}) = \frac{\sum_k E_{P_{im}}(k) E_{P_{jm}}(k)}{\sqrt{\sum_k |E_{P_{im}}(k)|^2} \sqrt{\sum_k |E_{P_{jm}}(k)|^2}}. \tag{8}$$

This is a measure with a range between 0 and 1, where 1 means that the two normalized envelopes are identical. Figure 3 shows this similarity measure for two different notes. In column A the similarity of the 25 first partials $(P_1, P_2, \ldots, P_{25})$ of a trumpet note are shown. The top graph shows the similarity of each partial relative to the first partial, $\beta(P_i, P_1)$. The bottom graph shows the similarity of each partial relative to its neighbor partial, $\beta(P_i, P_{i+1})$. Column B shows similar graphs for a flute note with vibrato. The amplitude modulations makes the partials less similar, but the general trend is the same. 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 on the note being played, but for many instruments, and harmonic instruments in particular, there is a significant correlation between the envelopes of the different partials of a single note.

D. Partial grouping

Each of the defined sensor partials contains one or more partials. In order to process these, the number of sources N and a mapping between the sources and the partial regions must be known. In other words, for each source it must be known in which partial regions its partials lie and for each partial region it must be known which sources that contribute a partial. Existing techniques can provide this information. In particular, for the examples in this paper, the number of sources N was obtained by applying clustering techniques in a spatial cue parameter space, as explained in [14]. The mapping between sources and partial regions was made based on harmonicity considerations [2].

Once this information has been established it is straightforward to find all the partial regions that contain only a single partial, as well as which source each partial is part of. For each source that has at least one non-overlapping partial, this provides a rough estimate of the temporal envelope shapes of all the other partials of that source. These rough estimates are called *model partials*, $Q_n(k,q)$, as they serve as models for the partials that need to be separated out of a partial region where they overlap. If all the partials of one source overlaps with partials of other sources the proposed method is not able to separate this source out of the mixture, unless additional information can be supplied.

E. Partial demixing

For each of the sensor partials containing overlapping partials from several sources, the mixing matrices $\mathbf{H}(q)$ (or its inverse) in (2) need to be estimated in order to recover the original source partials. This is achieved by an iterative search in the space of mixing matrices. For each candidate matrix, its inverse is applied to the sensor partials and the envelopes of the resulting "separated" partials are computed. The matrix that gives separated partials with envelopes whose shapes most closely resemble those of the model partials is chosen as the estimate of the mixing matrix for that corresponding partial region.

For most harmonic instruments the partials are narrow band. This means that the partial regions are also narrow band. Given the mixing model (2) it is possible to treat the different partial regions independently. The individual filters of the mixing matrix, $H_{mn}(q)$, depend on frequency. In other words, they vary over the frequency range of a given partial region. However, when the partial regions are narrow band, the filters change little over the actual frequency range. Therefore, for each partial region, the filters can be closely approximated by complex constants (filters with constant amplitude and constant phase). The separation problem is then equivalent to the problem of estimating the complex elements of a constant mixing matrix for each partial region Ω_i. For larger frequency ranges, it is possible to use filters with linear phase rather than constant phase. This does not increase the number of unknowns or the complexity of the problem. For notational simplicity only the former case is discussed in this paper.

Combining (2) and (4) gives

$$\begin{bmatrix} P_{i1}(k,q) \\ \vdots \\ P_{iM}(k,q) \end{bmatrix} = \begin{bmatrix} H_{11}(q) & \cdots & H_{1N}(q) \\ \vdots & \ddots & \vdots \\ H_{M1}(q) & \cdots & H_{MN}(q) \end{bmatrix} \begin{bmatrix} S_{i1}(k,q) \\ \vdots \\ S_{iN}(k,q) \end{bmatrix}, \tag{9}$$

where $S_{in}(k,q) = I_i(k,q)S_n(k,q)$ are the source partials. The mixing filters $H_{mn}(q)$ are different for the different partial regions (different frequencies), but are assumed to be constant within the coverage (frequency range) of the individual partial region. For a given partial region Ω_i, the mixing matrix does thus not depend on q, and is denoted $\mathbf{H}_i$. When the rank of the mixing matrix for a given partial region is equal to or greater than the number of partials in that partial region, separation of these partials is possible. This, in general, means that there must be as many sensors as there are overlapping partials in that region. However, the sources to which the overlapping partials belong to must be in different locations (giving M independent rows in the mixing matrix). For any estimate $\hat{\mathbf{H}}_i$ of the mixing matrix $\mathbf{H}_i$, separation is achieved by applying its (pseudo-)inverse $\hat{\mathbf{H}}_i^+$ on the known sensor partials. This is similar to frequency based blind source separation techniques. The difference lies in the way the mixing matrix is estimated.

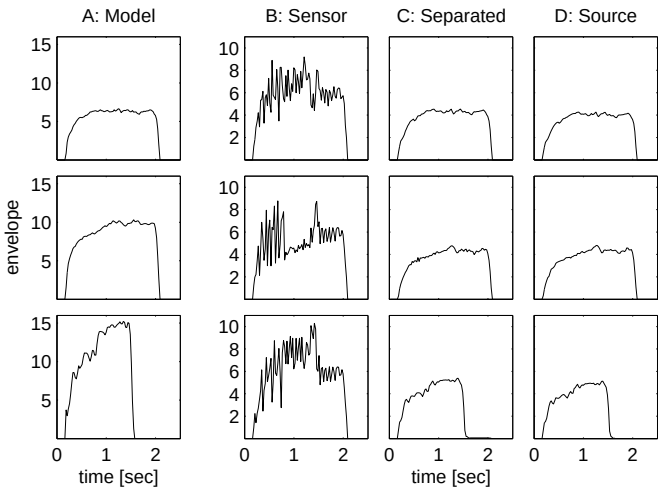

Fig. 4. *Partial envelopes for mixture of alto trombones, from left to right: A: model partials, A_5, C_4, and E_3 respectively (closest neighbors). B: sensor partials each containing overlapping A_6, C_5, and E_4. C: separated partials, A_6, C_5, and E_4 respectively. D: original source partials, shown for comparison.*

Left multiplying (9) with the estimated pseudo-inverse gives:

$$\begin{bmatrix} R_{i1}(k,q) \\ \vdots \\ R_{iN}(k,q) \end{bmatrix} = \hat{\mathbf{H}}_i^+ \begin{bmatrix} P_{i1}(k,q) \\ \vdots \\ P_{iM}(k,q) \end{bmatrix} = \hat{\mathbf{H}}_i^+ \mathbf{H}_i \begin{bmatrix} S_{i1}(k,q) \\ \vdots \\ S_{iN}(k,q) \end{bmatrix}, \tag{10}$$

where S_{in} are the source partials, and R_{in} are the *separated partials*. Each R_{in} represents the contribution of a single source S_n in the partial region Ω_i. It is obvious that under the given assumptions and with a correct estimate $\hat{\mathbf{H}}_i$ of the mixing matrix, the separated partials in (10) are identical to the source partials, i.e. perfect separation has been achieved.

For each partial region Ω_i the mixing matrix $\mathbf{H}_i$ is estimated by a search for the estimate $\hat{\mathbf{H}}_i$ that gives a best match between the separated partials R_{in} and the model partials Q_n. We achieve this by applying a standard multi-dimensional optimization technique (MATLAB `fminsearch`), maximizing the L_1 norm of the following similarity vector:

$$\beta_i = (\beta(R_{i1}, Q_1), \cdots, \beta(R_{iN}, Q_N)). \tag{11}$$

F. Practical considerations

1) Reducing complexity by disregarding reverberation: In general, the mixing matrix for a system with N sources and M sensors has $M \times N$ (complex) unknown elements. For a 3×3 system this gives an 18-dimensional space of candidate mixing matrices (9 complex elements).

A technique that can reduce the order of unknowns is to force all the elements of one row in the mixing matrix $\mathbf{H}$ to 1 [14], [15]. This can be done without loss of generality, as long as the real mixing filters H_{1n} contains no zeros. Then, (2) can be written:

$$\begin{bmatrix} X_1(k,q) \\ \vdots \\ X_M(k,q) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \frac{H_{21}}{H_{11}} & \cdots & \frac{H_{2N}}{H_{1N}} \\ \vdots & \ddots & \vdots \\ \frac{H_{M1}}{H_{11}} & \cdots & \frac{H_{MN}}{H_{1N}} \end{bmatrix} \begin{bmatrix} H_{11}S_1(k,q) \\ \vdots \\ H_{1N}S_N(k,q) \end{bmatrix}. \tag{12}$$

In the mixing matrix, the number of (complex) unknowns has been reduced by N. The formulation of the problem is the same as before (see e.g. (2)). The difference is that the real source signals $S_n(k,q)$, as emitted by the sources, has now been replaced by the same source signals as they would be received by the first sensor, $H_{1n}S_n(k,q)$. This means that no attempt is made to remove echoes or reverberation. However, in the context of source separation this can in general be accepted. The problems of echo cancellation, echo suppression, and dereverberation will not be discussed in this paper. It can be noted that other source separation techniques use STFT based techniques [14], [19], [20]. Our method can therefore easily be used as an extension to these existing methods.

2) Convergence: Even after applying the dimension reduction technique, the dimension of the space of mixing matrices remains high. When there are 3 sources and 3 sensors, the optimization algorithm trying to maximize (11) performs a search in a 12-dimensional space. The norm of the similarity vector (11) is a function in this space. Needless to say, it can take an arbitrarily complex form, and is not, in general, a concave function. The optimization algorithm may get trapped in some local maxima and not converge. In general, this depends on the starting estimate of the mixing matrix that is used in the iterative optimization algorithm. We have conducted several experiments on different sources and setups. In these experiments, although not being extensive, the choice of the L_1 norm in the optimization of 11 gave the best convergence behavior. We also experienced that when several sources have very similar temporal envelope shapes, this yields more local maxima. In this case the starting estimate is more critical for the convergence.

If the sensors and sources are in free field, each of the mixing filters will consist of a simple scaling factor and a pure delay. In this case it is possible to estimate these parameters from the model partials. For other partial regions (at different frequencies), the complex elements of the matrix $\hat{\mathbf{H}}_i$ are obtained by using the same scaling factors, but correcting the phases in order to yield the same delay. The pseudo-inverse $\hat{\mathbf{H}}_i^+$ can be computed directly, and the partials can be separated as in (10). Even though this is not feasible in real situations, the estimated free-field matrix $\hat{\mathbf{H}}_i$ can be used as the starting estimate in the iterative optimization algorithm.

In specific physical setups, it may be possible to further reduce the dimension of the problem. For instance, if the sensors are very closely spaced, the scale factor in the mixing filters can be approximated by the same constant for all the filters, leaving only the phases as unknowns. This effectively reduces the dimension by a factor of 2.

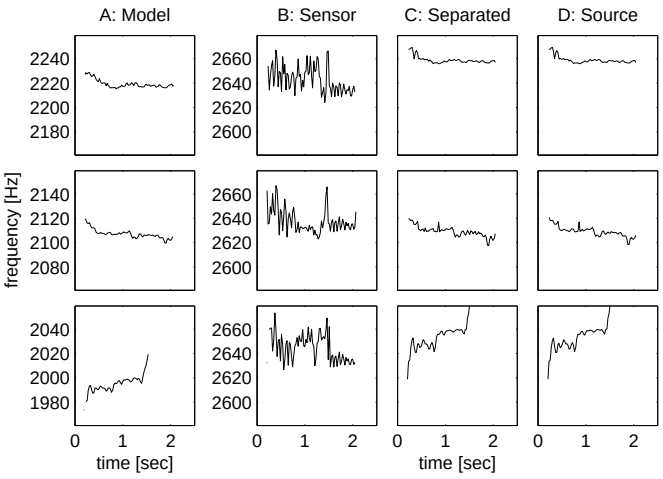

Fig. 5. *Partial frequency trajectories for mixture of alto trombones, from left to right: A: model partials, A_5, C_4, and E_3 respectively (closest neighbors). B: sensor partials each containing overlapping A_6, C_5, and E_4. C: separated partials, A_6, C_5, and E_4 respectively. D: original source partials, shown for comparison.*

3) Special case: When there are only 2 sensors (and maximum 2 simultaneously overlapping partials), each mixing matrix contains only two (complex) unknowns, since

$$\mathbf{H}_i = \begin{bmatrix} 1 & 1 \\ H_{21} & H_{22} \end{bmatrix}. \tag{13}$$

In this case, the inverse matrix is very simple:

$$\mathbf{H}_i^+ = \frac{1}{H_{22} - H_{21}} \begin{bmatrix} H_{22} & -1 \\ -H_{21} & 1 \end{bmatrix}. \tag{14}$$

Each of the rows of the demixing matrix has only one complex unknown, up to a scaling factor of $H_{22} - H_{21}$. Since our method is based on normalized envelopes, this scaling factor can be disregarded. This gives two independent equations, each with one unknown parameter. Thus, the original problem of dimension 4 has been split into 2 individual problems of dimension 2. This provides a fast computational method not involving any matrix inversions. The separation formula (10) consists of only 2 complex multiply-add operations for each time frame index k in the sensor partial envelopes. This allows for a more extensive, iteratively refined, search for the global maximum, in each of the two separate problems.

IV. Results

In order to demonstrate the performance of the presented method, a number of numerical simulations were carried out. In all the simulations, (2) was a good approximation of the mixing model and the method converged to a good estimate of the mixing matrix. Effectively, the overlapping partials were truly separated, recovering the amplitude modulation and frequency modulation of the original source partials. The quality of the separated sources therefore depends on how the residue is handled, i.e. the energy between the partials regions as well as the higher frequencies where no clear partials can be

detected. In an informal test we compared the DUET method [14] with a combined method using the proposed technique for different numbers of overlapping partials and DUET for the residue. Since DUET does not aim at truly separating the overlapping partials, it may introduce artificial beatings [8] and leakage between the separated sources. In the combined method the quality improved, even when the proposed method was applied only to one or two of the overlapping partials.

A. Three overlapping partials

The first example is a simulated situation of three sources and three sensors in free-field. The three notes A, C and E, all played by an alto trombone, were chosen as the source signals. These are the same notes as shown in Fig. 2. All the source signals were present at all the sensors. However, for each sensor, the source signals were scaled and delayed in order to simulate the wave propagation path. This setup provides us with three sensor signals. These sensor signals are quite similar since they all record the same scene. When the time-frequency analysis and grouping principles of section III are applied to either of these, the information shown in Fig. 2 is obtained. This provides us with a set of partial regions, as well as a mapping between these regions and the different sources (the partial labels that have been annotated in the figure).

Figure 4 shows the separation of the three overlapping partials in this signal. In column A the envelopes of the model partials are shown. Each of these model partials, $Q_n(k,q)$, corresponds to a different source, $S_n(k,q)$. In this example the closest non-overlapping partials were chosen as model partials, namely A_5, C_4, and E_3, respectively. Two of the model partials have quite similar envelopes, whereas the duration of the last model partial is shorter. These panels clearly show the smooth temporal envelopes of the original alto trombone notes, without beatings and strong amplitude modulations. In column B, the envelopes of the overlapping partials are shown for each of the three sensor signals. None of these are very regular or smooth since the three original partials interact and create heavy beatings, or amplitude modulations. Only toward the end, when one of the partials is silent (after about 1.5 sec), the envelopes are somewhat more regular. In column C, the envelopes of the separated partials are shown. These are A_6, C_5, and E_4, respectively. The three overlapping partials have been well separated, and the resemblance to the model partials is striking. In particular, we note that the beatings have disappeared and that the shorter partial (E_4) has been correctly recovered: after about 1.5 seconds the energy of the other two partials (longer duration) has vanished. Finally, in column D, the original (unknown) source partials are shown for comparison. The separated partials have accurately retained the amplitude of the original signals. Both the scale and the shape of the separated partials are more similar to the original source partials than they are to the model partials that were employed in the demixing algorithm.

Figure 5 shows the frequency trajectories for the same partials. The figure layout is the same as Fig. 4. In column A, the frequency trajectories of the model partials are shown. These are quite smooth, and relatively constant since the alto trombone exhibits no pronounced frequency modulation. If one disregards the start, where the attack transients affect the frequency estimates and the end, where the signal energy is very low, the frequency estimates are indeed almost constant. The ordinates also show how these model partials were chosen in different frequency regions. In column B, the frequency trajectories for the three sensor signals are shown. For each time frame the strongest spectral peak (interpolated) in the partial region was selected in order to form these trajectories. As previously mentioned, they are erroneous due to the fact that only one candidate peak can be chosen in each time frame and anyway all the peaks are the result of a superposition of overlapping partials. Consequently, the estimated frequency trajectories are noisy, as seen in the figure. Column C shows the frequencies of the separated partials. Qualitatively, these show the same behavior as the model partials. Both the constant frequency during the steady-state and the trend at onsets and offsets have been recovered. Finally, in column D, the frequency trajectories of the original source partials are shown. The trajectories of the separated partials accurately recover those of the original source partials.

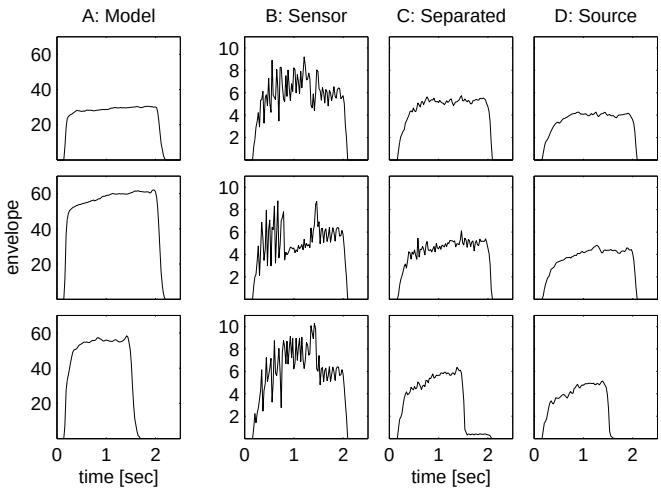

Fig. 6. *Partial envelopes for mixture of alto trombones, from left to right: A: model partials, A_1, C_1, and E_1 respectively (fundamental frequencies). B: sensor partials each containing overlapping A_6, C_5, and E_4. C: separated partials, A_6, C_5, and E_4 respectively. D: original source partials, shown for comparison.*

Effectively, three different partials whose frequency trajectories are less than 40 Hz apart and occasionally cross each other have been accurately separated from a 3 channel mixture.

We repeated the same separation example with another choice of model partials. When choosing model partials whose frequency bands lie farther away from the partial region containing overlapping partials, the envelopes of the model partials and the original source partials are in general less similar (see Fig. 3). In the new example the partial regions corresponding to the fundamental frequencies of the three sources, i.e. A_1, C_1, and E_1, were chosen as model partials. The separation result can be seen in Fig. 6. As in the first example, most of the beating has been removed. However, for the sources under analysis, the duration (offset) of the partials decreases with frequency, as can be seen in Fig. 1.

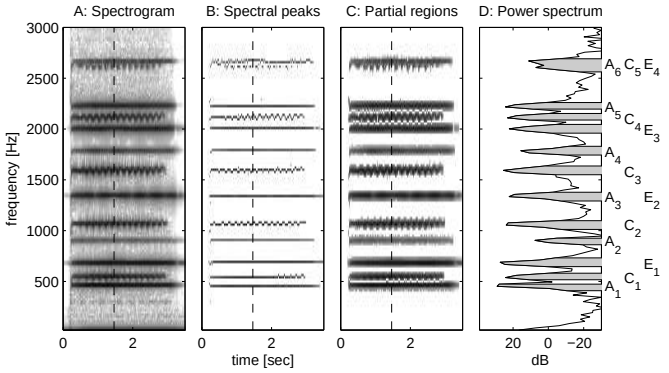

Fig. 7. *Three notes (A minor chord) played by violins, from left to right: A: spectrogram, B: spectral peaks, C: partial regions around the spectral peaks, D: power spectrum and detected partial regions for one particular time frame index (indicated by dashed lines in the other panels).*

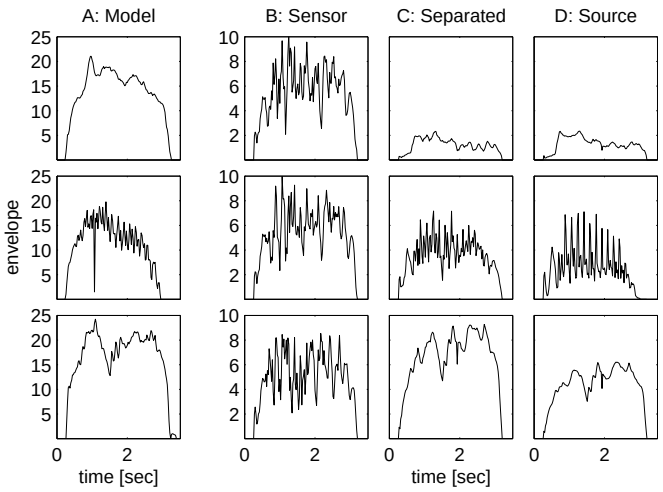

Fig. 8. *Partial envelopes for mixture of violins, from left to right: A: model partials, A_5, C_4, and E_3 respectively (closest neighbors). B: sensor partials each containing overlapping A_6, C_5, and E_4. C: separated partials, A_6, C_5, and E_4 respectively. D: original source partials, shown for comparison.*

This means that the chosen model partials slightly overestimate the duration of the overlapping partials. In this case, the estimate of the mixing matrix becomes slightly erroneous, resulting in some leakage between the separated partials. For example, some of the energy of the long duration A_6 and C_5 is still present in the separated E_4. This can be seen as a "tail" at the end of the envelope for the separated E_4. A possible explanation is that the tail increases the duration of this separated partial and yields higher similarity in (8) to the model partial whose duration was longer.

B. Frequency and amplitude modulation

A similar experiment, where the notes were played by a violin, is shown in Fig. 7-9. The notes A and E were played on open strings, without any pronounced frequency modulation. The C tone was played with vibrato, and the frequency modulations on its partials can be seen in Fig. 7. The vibrato also introduces amplitude modulations, as can be seen in the second row in Fig. 8. The amplitude modulation, although slightly different from that of the original source, is retained in the separated C_5. In the other two separated partials, which were without vibrato, the amplitude modulations have been correctly removed. Figure 9 shows the frequency trajectories for the partials. In the separated partials the frequency modulation of the vibrato has been retained in the C tone (with some error), whereas is has been removed in the other two partials.

When a note has strong amplitude modulations, the performance of the separation technique may degrade. This depends on the physics of the underlying instrument. For many instruments, frequency modulations are produced by varying the excitation, e.g. shaking the finger on the violin string board. In such cases, the frequency modulations of the different partials are synchronized (due to the change in effective length of the string imposed by finger movements). However, the amplitude modulations of the various partials are strongly related to the body modes of the instrument [25], that are, in general, not synchronized. This means that, even though different partial envelopes may look similar, their amplitude modulations may be out of phase, or even have different modulation frequencies. In such cases, the similarity measure in (8) is less meaningful. In the example above, the two notes without vibrato have led to correct estimates of the corresponding rows in the demixing matrix. For the note with vibrato, the estimate of the corresponding row in the demixing matrix is slightly erroneous because of the asynchrony in envelope modulations between the model and source partials. The difference in temporal envelope shape can be seen in Fig. 8. The erroneous row in the demixing matrix leads to scaling errors in the two other separated partials (correct shapes, but incorrect scale/phase) and errors in both the amplitude and frequency modulation of the separated vibrato partial.

It is possible to smooth the partial envelopes before computing the similarity in (8) in order to remove any strong amplitude modulations. In this case, however, most of the envelope information is lost and only the overall duration of the partials is significant in the separation. A better solution is to use similarity of frequency modulations as the similarity measure in the optimization method, as opposed to similarity of envelopes. For instruments where the frequency modulations of the different partials are in synchrony, like the violin, this can improve the separation quality. We have achieved this in some manual experiments. However, it is not clear how different similarity measures, namely similarity of envelopes and similarity of frequency modulations, should be combined in the similarity vector (11). In addition, there may be convergence issues, since the space in which this vector is to be maximized is rather complex.

C. Two overlapping partials

For a more realistic example, an artificial two-channel mixtures was generated by using head related impulse responses (HRIR) from the CIPIC database [26] as the mixing filters. These impulse responses are about 4.5 ms long filters (200 samples at 44.1kHz sampling rate), measured for a range of azimuths and elevations around various heads. Figures

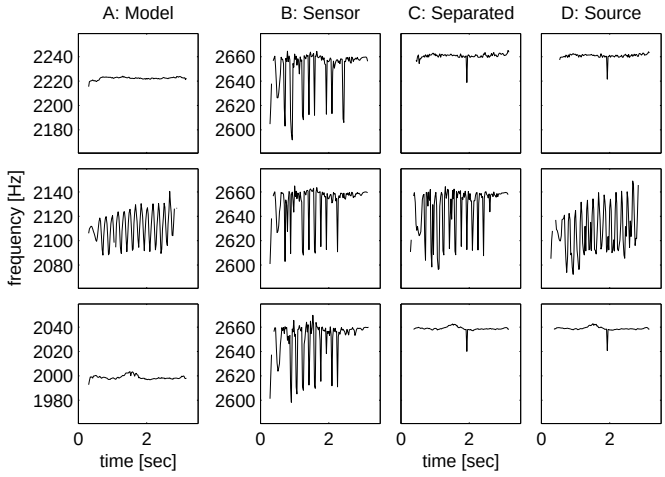

Fig. 9. *Partial frequency trajectories for mixture of violins, from left to right: A: model partials, A_5, C_4, and E_3 respectively (closest neighbors). B: sensor partials each containing overlapping A_6, C_5, and E_4. C: separated partials, A_6, C_5, and E_4 respectively. D: original source partials, shown for comparison.*

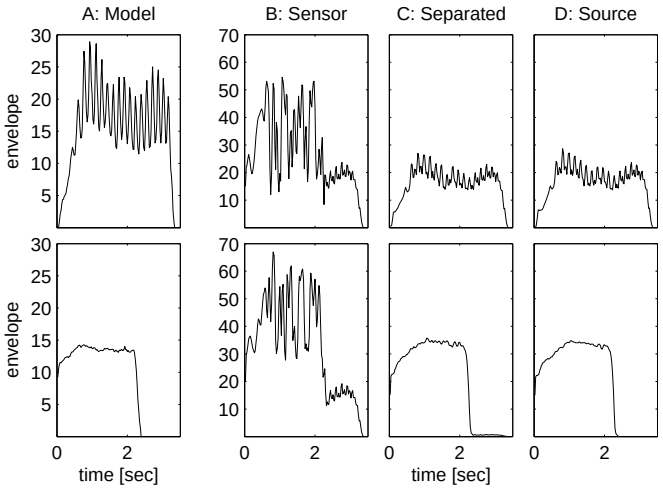

Fig. 10. *Partial envelopes, from left to right: A: model partials (closest neighbors). B: sensor partials. C: separated partials. D: original source partials.*

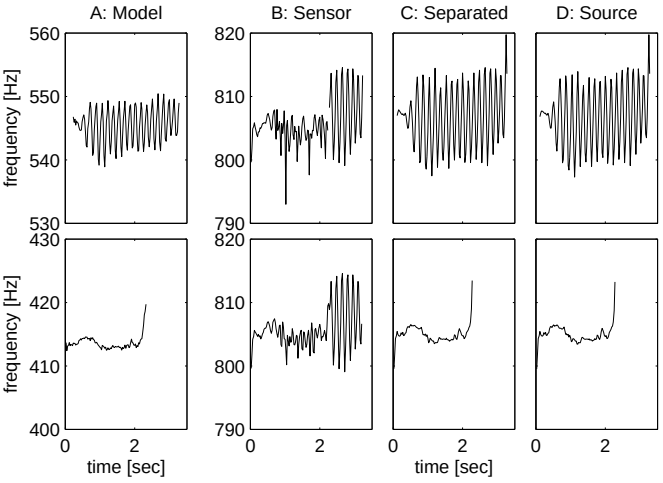

Fig. 11. *Partial frequency trajectories, from left to right: A: model partials (closest neighbors). B: sensor partials. C: separated partials. D: original source partials.*

10 and 11 show the envelope and frequency trajectories in the separation of two overlapping partials in a two-channel (binaural) mixture of two notes. One of the notes is played on a violin with vibrato, and the other note on a trumpet without vibrato. The figure shows the separation of the first overlap, i.e. the partial region (with overlapping partials) at lowest frequency. The closest neighbor partial regions were chosen as models. The figure layouts are the same as in previous examples, but with two panel rows in stead of three. The amplitude and frequency modulations of the violin are seen in the top panel rows in Fig. 10 and Fig. 11, respectively. In this case, for the violin partials containing vibrato, the amplitude modulations of the model partial and the original source partial are relatively synchronous and the envelope similarity measure gives nice separation. The separated partials have recovered the shapes of the original source partials to a high degree of accuracy, in both amplitude and in frequency.

We emphasize that the envelopes of the separated partials (column C) are better than what could be expected. Their shapes are closer to those of the perfect source partials (column D) than to those of the model partials (column A). Furthermore, even though the method works with normalized envelopes, the separated partials have retained not only the shapes of the original source partials, but also the correct scaling.

V. Conclusions

We have presented a method for separation of overlapping partials in multi-channel audio mixtures. The method combines aspects of time-frequency analysis and spatial demixing techniques. It is based on the maximization of the similarity of normalized envelopes of the partials and is able to accurately recover the amplitude and frequency modulation of the original source partials from the sensor signals where they overlap. The method works on partials individually, and can therefore also work (to some extent) in scenarios where there are more sources than sensors. Finally, it can easily be used in conjunction with several of the existing source separation methods.

Acknowledgments

The first author would like to thank Christof Faller for long discussions and constructive feedback during the preparation of this paper.

References

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

[2] 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*, Istanbul, Turkey, June 2000, pp. 765–768.

[3] ——, "Separation of harmonic sounds using linear models for the overtone series," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Orlando, Florida, USA, May 2002.
[4] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 56–69, January 1990.
[5] R. C. Maher, "Evaluation of a method for separating digitized duet signals," *Journal of the Audio Engineering Society*, vol. 38, no. 12, pp. 956–979, December 1990.
[6] T. Tolonen, "Methods for separation of harmonic sound sources using sinusoidal modeling," in *AES 106th Convention*, Munich, Germany, May 1999.
[7] M. Kazama, K. Yoshida, and M. Tohyama, "Signal representation including waveform envelope by clustered line-spectrum modeling," *Journal of the Audio Engineering Society*, vol. 51, no. 3, pp. 123–137, March 2003.
[8] H. Viste and G. Evangelista, "An extension for source separation techniques avoiding beats," in *Proceedings of 5th International Conference on Digital Audio Effects*, Hamburg, Germany, September 2002, pp. 71–75.
[9] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proceedings of 5th International Conference on Digital Audio Effects*, Hamburg, Germany, September 2002, pp. 51–58.
[10] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, August 1972.
[11] B. D. van Veen and K. M. Buckley, "Beamforming - a versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 4–24, April 1988.
[12] J. Peissig, "Binaurale hörgerätestrategien in komplexen störschallsituationen," Ph.D. dissertation, Universität Göttingen, Germany, 1993.
[13] T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. S. Woods, and B. Kollmeier, "Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction," *ACUSTICA united with acta acustica*, vol. 83, pp. 684–699, 1997.
[14] A. Jourjine, S. Rickard, and Ö. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Istanbul, Turkey, June 2000, pp. 2985–2988.
[15] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing (Elsevier)*, vol. 24, pp. 1–10, 1991.
[16] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Kyoto, Japan, September 4-6 1996, pp. 423–432.
[17] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
[18] T.-W. Lee, A. J. Bell, and R. H. Lambert, "Blind separation of delayed and convolved sources," *Advances in Neural Information Processing Systems, MIT Press*, 1997.
[19] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Int. Workshop on Independence and Artificial Neural Networks*, 1998.
[20] 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*, Aussois, France, January 1999.
[21] R. Plomp and W. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, pp. 548–560, 1965.
[22] A. S. Bregman, *Auditory Scene Analysis*. MIT Press, 1999.
[23] P. Fernandez-Cid and F. Casajus-Quiros, "Multi-pitch estimation for polyphonic musical signals," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 1998, pp. 3565–3568.
[24] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 708–716, November 2000.
[25] M. Mellody and G. H. Wakefield, "The time-frequency characteristics of violin vibrato: Modal distribution analysis and synthesis," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 598–611, January 2000.
[26] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, USA, October 2001, pp. 99–102.

LIST OF FIGURES