# MONAURAL SOUND SOURCE SEPARATION USING COVARIANCE PROFILE OF PARTIALS

*Priyank Goel, K R Ramakrishnan*

Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

## ABSTRACT

This paper addresses the problem of separation of pitched sounds in monaural recordings. We present a novel feature for the estimation of parameters of overlapping harmonics which considers the covariance of partials of pitched sounds. Sound templates are formed from the monophonic parts of the mixture recording. A match for every note is found among these templates on the basis of covariance profile of their harmonics. The matching template for the note provides the second order characteristics for the overlapped harmonics of the note. The algorithm is tested on the RWC music database instrument sounds. The results clearly show that the covariance characteristics can be used to reconstruct overlapping harmonics effectively.

***Index Terms***— monaural sound source separation, sinusoidal modeling

## 1. INTRODUCTION

Sound Source Separation has numerous applications in analysis, coding and manipulation of audio signals. A large set of sounds are harmonic sounds, which have special importance in music. Various efforts have been concentrated on the separation of harmonic sounds [1] [2] [3] [4]. The energy in the pitched sounds are concentrated in its harmonics and generally not all the harmonics overlap with the harmonics of the other concurrent sounds. Harmonics belonging to the same source have some similar properties. These properties can be exploited to reconstruct the overlapping harmonics with the help of non-overlapping harmonics.

The assumption that the rough shape of the amplitude spectrum of natural sounds is usually slowly-varying with respect to time and frequency, is known as spectral smoothness principle [5]. But this assumption is often violated in real instrument sounds [4]. The phenomenon that the amplitude envelopes of different harmonics of the same source tend to be similar, is known as common amplitude modulation (CAM) [4]. But CAM deteriorates with difference in the amplitudes of harmonics. Also in the instruments like violin and flute the excitation (bow-string contact and mouth-tip contact respectively) can change within a note. This causes the harmonic structure to change[6]. Hence the energy in one harmonic may increase while it is decreasing in another harmonic.

Spectral Smoothness assumes a first order relation between partials, i.e. the amplitude of a partial depends directly on the amplitude of other partials. On the other hand CAM assumes second order relation of partials (the variation in the amplitude of a partial depends upon the variation in the amplitude of other partials). But rather than assuming non-uniform covariance across harmonics, it oversimplifies by assuming uniform covariance between harmonics. To overcome the limitations of the above two assumptions in dealing with real world sounds, we propose to use the full covariance profile to reconstruct the overlapping harmonics.

Sinusoidal modeling has been used to decompose audio signals into their deterministic or sinusoidal and stochastic parts. Since we have used only harmonic sources we assume that all the sources can be faithfully represented by sinusoidal modeling. Following this assumption, we have used sinusoidal modeling for representations and calculations in our work. Our aim in this paper is to prove the relevance of the new feature in reconstructing overlapping harmonics rather than presenting an end-to-end source separation system. Hence we assume that the onset, offset, pitch and source instrument of each note present in the mixture are known.

## 2. THE PROBLEM FORMULATION

The time axis is divided into frames and the whole processing is done frame-wise. Suppose a note has $H$ harmonics and its duration is $L$ frames. Then the sinusoidally modeled part of the signal at $l^{th}$ frame of a note is given as:

$$s_l(n) = \sum_{h=1}^{H} a_{h,l} \cos(2\pi f_{h,l} n + \phi_{h,l}), \qquad n = 1, \ldots, N \tag{1}$$

Here $a_{h,l}$, $f_{h,l}$ and $\phi_{h,l}$ are the amplitude, frequency and initial phase of the $h^{th}$ sinusoid respectively, at the $l^{th}$ frame of the note. Now, in a given musical piece we have $J$ such notes starting at their respective onset frames $o^j$. The number of frames in note $j$ is given by $L^j$. Using note lengths ($L^j$) and onsets($o^j$) we can find out which notes are present at a particular frame $m$ of the music piece. Let $\mathcal{J}_m$ be the set of notes

© EURASIP, 2012 - ISSN 2076-1465

present in a given frame $m$. Then

$$\mathcal{J}_m = \{j | o^j \leq m \leq o^j + L^j + 1, \forall j \leq J\} \quad (2)$$

So the sinusoidally modeled part of the musical piece in a given frame $m$ can be represented as:

$$\hat{x}_m(n) = \sum_{j \in \mathcal{J}_m} \sum_{h=1}^{H^j} a^j_{h,m-o^j+1} \cos(2\pi f_{h,m-o^j+1} n + \phi_{h,m-o^j+1}) \quad (3)$$

At a given frame $m$ if cardinality $|\mathcal{J}_m| > 1$, then more than one note exist at that frame. Since we know the pitch of the notes, we can estimate the frequency of the harmonics of the notes. Since most of the energy of the notes is concentrated at these frequencies, if harmonics of two different notes are far apart in frequency they can be faithfully reconstructed. To decide which harmonics are close in frequency or overlapping with harmonics of other note, a threshold $f_t$ is used. So if two sources $j_1, j_2$ occur at frame $m$, i.e. $j_1, j_2 \in \mathcal{J}_m$, we say harmonic $h_1$ of source $j_1$ and harmonic $h_2$ of source $j_2$ are overlapping if

$$|f_{h_1,m-o^{j_1}+1} - f_{h_2,m-o^{j_2}+1}| < f_t \quad (4)$$

A simple least square formulation can be used to estimate the parameters of non-overlapping harmonics but not of overlapping harmonics. In this paper we address the problem of finding the parameters of overlapping harmonics. After solving this it is straightforward to reconstruct the notes.

Consider the matrices $\mathbf{A}^j$, $\mathbf{F}^j$ and $\mathbf{\Phi}^j$ for each note $j$, which will store respectively the amplitudes, frequencies and initial phases of the sinusoids of the note. The element of the matrix $\mathbf{A}^j$ at row $h$ and column $l$, i.e. $a_{h,l}$ will give the amplitude of $h^{th}$ harmonic at $l^{th}$ frame. Let $\mathcal{L}^j$ and $\mathcal{H}^j$ be the set of frames and set of harmonics for note $j$. We define a set $\mathcal{C}^j$ to store the indices of the overlapping or corrupted elements of $\mathbf{A}^j$ or $\mathbf{\Phi}^j$.

$$\begin{aligned}\mathcal{C}^j = \{(h,l) | \, |f_{h,l+o^j-1} - f_{\acute{h},l+o^j-o^i}| < f_t, \\ j \neq i, i \in \mathcal{J}_{l+o^j-1}, l \in \mathcal{L}^j, h \in \mathcal{H}^j, \acute{h} \in \mathcal{H}^i\}\end{aligned} \quad (5)$$

A set $\mathcal{U}^j$ is similarly defined to store the indices of uncorrupted harmonics. We also define a set $\mathcal{H}^{j_u}$ containing the harmonics which are uncorrupted throughout the occurrence of the note.

$$\mathcal{H}^{j_u} = \{h | (h,l) \in \mathcal{U}^j, \forall l \in \mathcal{L}^j\} \quad (6)$$

These will also be called totally uncorrupted harmonics. A similar set $\mathcal{H}^{j_c}$ will contain the harmonics which are corrupted at least for one frame. Our aim is to estimate the correct values for matrices $\mathbf{A}^j s$ and $\mathbf{\Phi}^j s$ at indices given by $\mathcal{C}^j s$. First amplitudes of sinusoids at overlap regions are estimated. After that the corresponding phase values are calculated.

## 3. AMPLITUDE ESTIMATION

This section presents the method for estimating the amplitude of sinusoids at overlapping regions. The estimation for each note is done separately and independent of other notes. Rather than working with amplitudes of the sinusoids, we use their log amplitudes.

$$\mathbf{B}^j = \log(\mathbf{A}^j) \quad (7)$$

Here the logarithm is element wise. A matrix $\mathbf{B}^{j_u}$ is formed which contains the log amplitudes of totally uncorrupted harmonics. Each column of this matrix is treated as a feature vector. The sample mean of each row of matrix $\mathbf{B}^{j_u}$ is subtracted from its elements to get the zero mean matrix $\mathbf{Z}^{j_u}$. We then calculate the covariance matrix $\mathbf{\Sigma}^{j_u} = [\sigma_{h1,h2}]_{Hj_u \times Hj_u}$ as the sum of outer products

$$\sigma_{h1,h2} = \frac{1}{L^j} (\mathbf{z}'_{h1} \cdot \mathbf{z}_{h2}) \quad (8)$$

Where $\mathbf{z}_{hn}$ is the $n^{th}$ row of matrix $\mathbf{Z}^{j_u}$. The next task is to find a template sound which matches the covariance matrix of the note $j$ calculated above.

A template sound is actually a part of or a full note which does not coincide with any other note. We look for the continuous regions in the given musical piece where $|\mathcal{J}_m| = 1$. Suppose there are $I$ such regions and in such a region $i$, a note $j$ exists. A log amplitude matrix $\acute{\mathbf{T}}^i$ is calculated using part of note $j$ in region $i$. We subtract the mean of each row of matrix $\acute{\mathbf{T}}^i$ from its elements to calculate the zero mean matrix $\mathbf{T}^i$. This $\mathbf{T}^i$ is our template matrix. We get $I$ such template matrices by this procedure.

After getting the templates we will now do the matching. For a given note $j$, we try to find the best matching template, independent of other notes. The first necessary condition to be satisfied by a template $i$ to match is that the number of harmonics in the template should be more than number of harmonics of the note $j$. If the above condition is satisfied, we try to match the covariance profile of the template and the note. To do that, we first construct a matrix $\mathbf{T}^{ij_u}$ containing the rows of $\mathbf{t}_h$ corresponding to uncorrupted harmonics of note $j$. We do Principal Component Analysis of $\mathbf{T}^{ij_u}$ to get principal vectors matrix $P^{ij_u}$. Now we check whether the principal vectors $\mathbf{P}^{ij_u}$ can diagonalize the covariance matrix $\mathbf{\Sigma}^{j_u}$ as:

$$\mathbf{D}^{ij_u} = (\mathbf{P}^{ij_u})' \mathbf{\Sigma}^{j_u} \mathbf{P}^{ij_u} \quad (9)$$

Here $\mathbf{D}^{ij_u}$ is the resultant matrix of the diagonalization. To quantify the amount of diagonalization we define the measure:

$$\gamma^{ij} = \left( \sum_{m=n} (d^{ij}_{m,n})^2 \right) - \left( \sum_{m \neq n} (d^{ij}_{m,n})^2 \right) \quad (10)$$

If the principal vectors $\mathbf{P}^{ij_u}$ diagonalize the covariance matrix $\mathbf{\Sigma}^{j_u}$ well, then this means that the covariance profile of

template $i$ matches with note $j$. A template for the note $j$ for which $\gamma^{ij}$ is maximum is chosen for its reconstruction.

The reconstruction is obtained using Principal Component Analysis (PCA). We do PCA of matrix $\mathbf{Z}^{j_u}$ to find out the principal vector matrix $\mathbf{P}^{j_u}$ and projection matrix $\mathbf{Y}^{j_u}$.

$$\mathbf{Z}^{j_u} = \mathbf{P}^{j_u}\mathbf{Y}^{j_u} \tag{11}$$

We have $H^{j_u}$ principal vectors to represent uncorrupted harmonics of note $j$. For each such vector, we will find out a corresponding vector to reconstruct corrupted harmonics of this note. First we will construct a matrix $\mathbf{T}^{ij}$ by truncating the columns of $\mathbf{T}^i$ to contain the same number of harmonics(or rows) as $j$. Now let $\mathbf{p}_m^u$ be the $m^{th}$ principal component , i.e. the $m^{th}$ column of matrix $\mathbf{P}^{j_u}$. We project $\mathbf{T}^{ij}$ on a lower dimensional space in which all the dimensions corresponding to $h \in \mathcal{H}^{j_u}$ are replaced by one vector $\mathbf{p}_m^u$. We calculate only the first principal component of this projected data. Only the first coefficient in this principal vector correspond to the direction of $\mathbf{p}_m^u$ and others correspond to the basis $h \in \mathcal{H}^{j_c}$. After removing that coefficient we get the $m^{th}$ principal vector $\mathbf{p}_m^c$ for corrupted harmonics. We calculate the other principal components in the same way and gather them as columns of matrix $\mathbf{P}^{j_c}$.

The matrix $\mathbf{Y}^{j_u}$ contain the weights of the vectors in $\mathbf{P}^{j_u}$ to reconstruct the zero mean log amplitude vectors of uncorrupted harmonics. $\mathbf{P}^{j_c}$ contains the vectors for corrupted harmonics corresponding to each vector in $\mathbf{P}^{j_u}$. So we calculate the zero mean log amplitude matrix for corrupted harmonics as:

$$\mathbf{Z}^{j_c} = \mathbf{P}^{j_c}\mathbf{Y}^{j_u} \tag{12}$$

These zero mean log amplitude values are then utilized for interpolation of the harmonics in the overlap regions. The interpolation is done one harmonic at a time using the non-overlapping log amplitude values which are calculated by least square estimation and zero mean log amplitudes calculated by our proposed method. There are scenarios where a harmonic is corrupted throughout the note. In these cases interpolation is not possible. We calculate the mean log amplitude of such harmonic by interpolation from mean log amplitudes of nearby harmonics using spectral smoothness assumption. Please note that we are assuming smoothness of the mean values of log amplitudes of the harmonics rather than all the log amplitude values of the harmonics. We add this mean to our calculated zero mean values of the harmonic to get the log amplitudes. The log amplitudes are converted to amplitudes using exponential operation.

The feature proposed by us is of use only if at-least two harmonics of the note to be reconstructed are totally uncorrupted. Though this is a realistic assumption in most cases, it violates when octaves are being played in music. The presented method is extensible to any number of sources present, and the separation quality will only be affected by the amount of overlap of harmonics and independent of number of sources. In the next section we describe the reconstruction of phases using these amplitudes. Although the method is described for presence of two sources at a time, it can easily be extended to 3 or more sources.

## 4. PHASE RECONSTRUCTION

We now have the amplitudes of all the sinusoids. We will now derive phases of sinusoids in the overlap regions. We have to take the following two things into consideration:

1. There should be continuity between the sinusoids of consecutive frames.
2. The error between observed mixture and reconstructed notes has to be minimized.

While estimating the amplitudes at overlaps, we have taken care for retaining continuity by using interpolation. For initial phases at frames, we must take care that the initial phase of the sinusoid at the present frame should match its phase from the previous frame to get a continuous sinusoid across the two frames. Suppose $i$ is the current frame and the frequency and initial phase of the previous frames be $f_{i-1}$ and $\phi_{i-1}$ respectively. If the time shift between the frames is $T_s$. To maintain continuity of the harmonic, the initial phase of the current frame is given by:

$$\hat{\phi}_i = \pmod{f_{i-1}T_s + \phi_{i-1}, 2\pi} \tag{13}$$

To maintain continuity, the value of $\phi_i$ should be nearby $\hat{\phi}_i$. We set a threshold $\delta_t$ for $\delta_\phi = |\phi_i - \hat{\phi}_i|$ such that when,

$$\delta_\phi < \delta_t \tag{14}$$

we say that $\phi_i$ is continuous. While estimating the initial phases at overlap regions, we will take care to maintain continuity by above stated relation.

The relation given in (13) is for ideal case. In practical scenario there are lot of factors which cause deviation of observed phase with that given by (13). So if extrapolation of the phase values in overlap regions is done just using the relation (13) then there will be accumulation of reconstruction error with the length of extrapolation. Hence it would be better to take help from the observed mixture to find the phase values. The phase values which reduce the difference between observed and reconstructed data within the constraint given by (14) are desired.

We now explain the procedure of phase reconstruction in detail. We first figure out the frames for which $|\mathcal{J}_m| > 1$. Let $m$ be such a frame. Let $j_1$ and $j_2$ be two sources which exist at this frame, i.e. $\mathcal{J}_m = \{j_1, j_2\}$. Let $l_1$ and $l_2$ be the corresponding column numbers for frame $m$ in matrices $\mathbf{\Phi}^{j_1}$ and $\mathbf{\Phi}^{j_2}$. Suppose harmonic $h_1$ of note $j_1$ overlaps with harmonic $h_2$ of note $j_2$. We will first find out the constraint for continuity for the phase of harmonic $h_1$. The same procedure will be applied on $h_2$ too. There are two possible cases:

1. The phase of the harmonic is known at either side of the frame, i.e. $\phi_{h_1,l_1-1}$ or $\phi_{h_1,l_1+1}$ is known.

2. The phase of the harmonic is unknown at both sides, i.e. both $\phi_{h_1,l_1-1}$ and $\phi_{h_1,l_1+1}$ are unknown.

For the first case we can find out the expected phase $\hat{\phi}_{h_1,l_1}$ with the use of (13).

The threshold mentioned in (14) is actually frequency dependent. We define a constant $t_t$(typically $t_t = 0.0005$) whose unit is time.The threshold $\delta_{t_1}$ is calculated as:

$$\delta_{t_1} = \pi t_t f_{h_1,l_1} \tag{15}$$

When $\delta_{t_1} > \pi$, we set $\delta_{t_1} = \pi$. The above relation is based on the fact that the perceptual sensitivity to the continuity of phase decreases as frequency increases [7]. The constraint on $\phi_{h_1,l_1}$ to ensure continuity is:

$$|\hat{\phi}_{h_1,l_1} - \phi_{h_1,l_1}| \leq \delta_{t_1} \tag{16}$$

For the second case when both $\phi_{h_1,l_1-1}$ and $\phi_{h_1,l_1+1}$ are unknown, we do not have any information and hence no constraint for continuity too. So we take $\hat{\phi}_{h_1,l_1} = 0$ and $\delta_{t_1} = \pi$.

The above mentioned procedure is applied also for $\phi_{h_2,l_2}$ to get $\hat{\phi}_{h_2,l_2}$ and $\delta_{t_2}$. Now the range in which $\phi_{h_1,l_1}$ and $\phi_{h_2,l_2}$ should be located to maintain continuity is known.

The next step is the search for optimum $\phi_{h_1,l_1}$ and $\phi_{h_2,l_2}$ which will minimize the difference between the observed data $x_m$ and sinusoids produced by these parameters.

$$\begin{aligned}\xi =&|x_m(n) - (a_{h_1,l_1}\cos(2\pi f_{h_1,l_1}n + \phi_{h_1,l_1}) \\ &+ a_{h_2,l_2}\cos(2\pi f_{h_2,l_2}n + \phi_{h_2,l_2}))|\end{aligned} \tag{17}$$

Our aim is to minimize $\xi$. We do this by declaring some candidates for $\phi_{h_1,l_1}$ and $\phi_{h_2,l_2}$ which follow the constraint derived and choosing the pair of candidates which minimize $\xi$. The above procedure is repeated for all the overlap regions to get all the phase values.

## 5. EXPERIMENTAL RESULTS

Simulation experiments were carried out to evaluate the performance of the proposed method. The sound signals were taken from RWC Musical Instrument Sound Database [8]. The database for experiment consisted of 4 pieces from Pianoforte, 1 piece from Harmonica (Blues Harp), 6 pieces from Classic Guitar (Nylon String), 1 piece from Ukulele, 4 pieces from Mandolin, 4 pieces from Violin and 2 pieces from Flute. The Musical pieces were of length $30s$ each. A musical piece from one instrument was added with a musical piece from another instrument to get a mixture. 63 such mixtures of duration $30s$ each were prepared for simulation experiments. The onset, offset and pitch values of notes were calculated from the pieces before mixing.

The sampling rate used was 16 kHz and frame size was 1024 samples. The signal to noise ratio of the estimated signal is given by:

$$SNR_{est} = 10\log\frac{\sum_n x_o^2(n)}{\sum_n (x_o(n) - x_e(n))^2} \tag{18}$$

The signal to noise ratio of the mix is given by:

$$SNR_{mix} = 10\log\frac{\sum_n x_o^2(n)}{\sum_n (x_o(n) - x(n))^2} \tag{19}$$

Here, $x_o$ is the original source signal prior to mixing, $x_e$ is the estimated source signal from the mixture and $x$ is the mixture signal. The $SNR$ gain is then $\Delta SNR = SNR_{est} - SNR_{mix}$. The mean value of the total gain in $SNR$ from all the mixtures was 18.659. The minimum gain in $SNR$ was $-2.8930$ and maximum was 36.1355. A histogram of total gain in $SNR$ from the simulations is presented in figure 1.

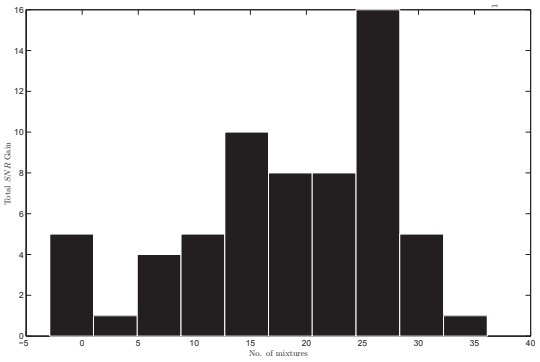


**Fig. 1**. Histogram showing the total $SNR$ gain obtained by applying the proposed algorithm on 63 mixtures.

Our algorithm performs better than many of the state-of-art sound separation algorithms. The algorithms by Li[4] and Virtanen[5] give 14.7 and 11.0 gain in $SNR$ respectively while using ground truth pitch. Our algorithm gives negative $SNR$ improvement for 8% cases. This is because of following reasons:

- **Inharmonicity of partials -** The frequencies of harmonics was estimated using the pitch values. Though our algorithm takes care of inharmonicity to some extent, in many cases the estimation can go wrong. Sometimes a harmonic from a source can get be wrongly attributed to some other source. This increases the error by a lot.

- **Estimation of amplitudes without considering observed data -** The amplitudes at overlap regions were estimated by just using covariance profile of the note and not the observed data at those regions. This has at times resulted in amplitudes which do not accede well with the observed data.

- **Effect of overlap on regions considered as non-overlapping -** Though we have set a threshold($f_t$) on the difference between frequencies of harmonics to decide whether or not they are overlapping, some energy of the harmonics is spread farther than $f_t$. These cause error in estimation of amplitudes and phases of elements which are considered uncorrupted. Interpolation/extrapolation using these erroneous values sometimes leads to unexpected amplitude trajectories.

In spite of these limitations, the overall performance of our algorithm is very good. The strength of this algorithm lies in its ability to separate highly dynamic sounds. Examples of separated signals are available at {http://iris.ee.iisc.ernet.in/index_files/priyank.htm}.

## 6. CONCLUSIONS

A new feature for reconstruction of amplitudes of overlapping harmonics in musical recordings is presented. Also a constrained search for phase reconstruction is proposed with frequency dependent constraint. Our method performs well even for instruments with changing harmonic structures such as violin and flute. The method requires onset,offset and pitch values a priori, hence can be implemented along with a multi-pitch estimator.

## 7. REFERENCES


[1] Tuomas Virtanen and Anssi Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, may 2002, vol. 2, pp. II–1757 –II–1760.

[2] P. Jinachitra, "Constrained em estimates for harmonic source separation," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, 2003, vol. 2, pp. II – 617–20 vol.2.

[3] Zhiyao Duan, Yungang Zhang, Changshui Zhang, and Zhenwei Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 766 –778, may 2008.

[4] Yipeng Li, J. Woodruff, and DeLiang Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1361 –1371, 2009.

[5] Tuomas Virtanen, *Sound Source Separation in Monaural Music Signals*, Ph.D. thesis, Tampere University of Technology, Finland, November 2006.

[6] Arnold Small, "The harmonic structure of the violin tone as a function of bow speed and bow pressure," *The Journal of the Acoustical Society of America*, vol. 9, no. 1, pp. 79–79, 1937.

[7] Albert S. Bregman, "Computational auditory scene analysis," chapter Psychological data and computational ASA, pp. 1–12. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1998.

[8] Masataka Goto, "Development of the rwc music database," in *Proceedings of the 18 th International Congress on Acoustics (ICA 2004)*, 2004, pp. 553–556.