Methods for Separation of Harmonic Sound Sources using Sinusoidal Modeling

Tero Tolonen
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Espoo, Finland
Tero.Tolonen@hut.fi
http://www.acoustics.hut.fi/~ttolonen

Abstract

Methods are proposed for separation of harmonic sound sources using sinusoidal modeling. A local nonlinear least-squares (NLS) frequency estimator is proposed to resolve sinusoids that are close in frequency. An iterative analysis scheme using interpolated parameter trajectories and subtraction of detected components is presented. A measure is proposed for testing the accuracy of the model.

0 Introduction

Separation of sound source signals is a task that is often encountered in digital audio. Historically, it has been applied mainly in two areas: separation of speech of two or more simultaneous talkers and segmentation and transcription of musical signals. The currently active research fields of sound source separation include compression coding of audio signals, computational auditory scene analysis, automatic transcription of music, and analysis/synthesis applications in computer music.

Separation of sound sources in a mixed sound signal is a difficult problem and no reliable methods are available for the general case. However, when certain assumptions can be made on the properties of the sound sources or the signals they produce, a mixed signal consisting of several contributing sources may be separated into signals that are perceptually close to the original signals before mixing. In this study, we concentrate on mixtures of harmonic or nearly harmonic musical signals and their separation.

This contribution proposes three methods for improving the accuracy of sound source separation based on sinusoidal modeling: 1) a local nonlinear least-squares (NLS) estimator for improving the frequency resolution in detection of parameters of sinusoids that are

close in frequency, 2) an iterative analysis scheme using interpolated parameter trajectories and subtraction of detected components, and 3) a measure for testing the accuracy of the model.

The paper is organized as follows. Section 1 reviews sinusoidal modeling techniques and presents an outline of the sound separation algorithm. In section 2, an iterative sinusoidal modeling algorithm with interpolated parameter trajectories is proposed that is well suited for separation of harmonic sound sources. In addition, a measure is proposed that can be used for estimating the relevance of each sinusoid. Section 3 describes a local nonlinear least-squares (NLS) method for resolving parameters of sinusoids that are not resolved by the discrete Fourier transform (DFT). Performance of the method is demonstrated in Section 4 with a sound separation example. Finally, Section 5 concludes the paper and proposes research paths for future developments.

1 Sound Separation based on Sinusoidal Modeling

One of the promising methods in sound source separation is sinusoidal modeling that is based on the short-time Fourier transform (STFT) (McAulay and Quatieri, 1986; Smith and Serra, 1987; Serra, 1989; Serra and Smith, 1990). It is suited for separation of sound sources in musical signals since many musical instruments produce harmonic or nearly harmonic signals with relatively slowly varying sinusoidal partials. Sinusoidal modeling offers a parametric representation of these signal components such that the original signal can be recovered by synthesis and addition of the components. By grouping the sinusoidal components to correspond to harmonic tones, the essential features of the sound source contributions may be separated. Note that since few musical instruments produce strictly periodic signals, other methods for representing noise and transient components may have to be used as well. At this time it appears a good compromise offering generality and relatively good analysis accuracy.

1.1 Overview of Sinusoidal Modeling Techniques

Sinusoidal modeling is a set of techniques in which a sound signal is represented as a set of sinusoids that are parameterized by amplitude, frequency, and phase trajectories (McAulay and Quatieri, 1986; Smith and Serra, 1987). Later, the sinusoidal modeling has been accompanied with models for noise (Serra, 1989; Serra and Smith, 1990) and for transients and noise (Verma et al., 1997; Ali, 1995; Hamdy et al., 1996; Levine, 1998).

Numerous modifications and elaborations have been proposed for different sinusoidal modeling applications. Iterative sinusoidal analysis algorithms have been presented in (George and Smith, 1992, 1997; Macon, 1996; Edler et al., 1996; Edler and Purnhagen, 1998). Heterodyne filtering has also been applied in parameter estimation (Ding and Qian, 1997). Analysis schemes using different time resolution in different frequency bands have been presented in (Rodríguez-Hernández and Casajús-Quirós, 1994; Anderson, 1996; Levine et al., 1997, 1998; Levine and Smith, 1998; Goodwin, 1997).

Depalle and Tromp (1996) and Depalle and Hélie (1997) propose a method for improving the estimation of frequencies, amplitudes, and phases of the sinusoidal components. The method allows for reduction of the length of the analysis window to 1.5 periods of the lowest frequency of interest. Prandoni et al. (1997) propose a method for window size optimization using dynamic programming. A technique for synchronizing the analysis windows of sinusoidal modeling to the transient events in the analyzed signal is presented by Masri and Bateman (1996).

A method of signal component parameter trajectories estimation based on *hidden Markov models* (HMM) is presented by Depalle et al. (1993). The method is promising since it is reported that the parameters of two crossing partials can be determined using the method. In general, the peak continuation algorithm of sinusoidal modeling is one of the most crucial parts of the analysis system. Similar approaches for STFT parameter estimation using HMMs are described by Streit and Barrett (1990) and Barrett and Holdsworth (1993).

The above overview of sinusoidal modeling techniques is by no means complete. Its purpose is just to get a brief review of some of different sinusoidal analysis methods presented in the literature. In the following, we discuss sinusoidal modeling in the context of separation of harmonic sound sources.

1.2 Sinusoidal Modeling in Separation of Harmonic Sources

Figure 1 presents a general block diagram of sound source separation based on sinusoidal modeling. Typically, a pre-analysis block is included to provide information for separation of the sound source signals. One of the tasks of pre-analysis is to locate sound events and it may include, e.g., pitch analysis and onset and offset detection. Pre-analysis may provide a basis on which the sinusoidal analysis parameters, such as the window size and the hop size, are chosen. Its results are also used in grouping the sinusoidal components. For instance, pitch analysis results can be used in selecting sinusoids corresponding to a harmonic tone.

The sinusoidal modeling block is the core of the separation system providing a low level representation of the signal as sinusoids. The sinusoidal modeling block may also include noise and transient modeling functionalities in which case the low level representational elements are sinusoids, noise components, and transient components.

After the signal has been decomposed into low level signal elements, a grouping algorithm is used to form sets of the elements corresponding to sound events. In the broad sense the grouping block may also provide association of the events into event streams each of which corresponds to a single sound source. However, the stream formation is beyond the scope this paper.

Figure 1 also provides a feedback path in which a prominent separated and synthesized sound source signal may be removed from the original signal for more accurate analysis of the remaining sound source components. A sound separation example presented in

Section 4 shows how the iterative sound source separation may be used to detect weaker tones that are not observable directly.

It is clear that sinusoidal modeling alone is not a sufficient representation for audio signals in general and that methods for noise and transient components are required in a sound separation system. However, in this contribution we concentrate on developing techniques that improve the accuracy and quality of sinusoidal modeling. We assume that this will also improve the performance of noise and transient modeling methods.

2 Iterative Analysis

In iterative sinusoidal modeling the analysis stage is a recursive algorithm which in each step detects and removes the most prominent sinusoidal component. It thus automatically sorts the sinusoidal parameters according to the prominence criterion used. Iterative sinusoidal modeling is computationally considerably more expensive than the regular non-recursive analysis but it may provide substantially more accurate parameter estimation, as shown in an example below.

Figure 2 shows a general block diagram of iterative sinusoidal analysis. Analysis algorithms used in parametric audio coding (Edler et al., 1996), analysis and synthesis of musical tones (George and Smith, 1992) and of speech (George and Smith, 1997; Macon, 1996) are examples of this approach. The input to the recursive algorithm is a windowed segment of an audio signal. In each iteration, the most prominent sinusoid is first detected. In (George and Smith, 1992, 1997), the prominence criterion is the energy of the residual signal, i.e., the aim is to select a sinusoid which minimizes the energy of the residual signal that is obtained by subtracting the synthesized sinusoid from the original one. In (Edler et al., 1996), on the other hand, the criterion is based on a psychoacoustic model which attempts to select the perceptually most significant component.

After the most prominent component is detected, its parameters are estimated. Using the estimated parameters, a representation of the sinusoid is synthesized and removed from the previous residual signal or the original signal in the first iteration. The removal is typically performed by subtracting the synthesized sinusoid from the residual of the previous step in the time or the frequency domain, depending on the parameterization. Naturally, if time-domain subtraction is used, a new DFT representation is required in each recursion. The recursion is continued until all significant components have been detected and removed. The significance criteria are discussed in the following subsection.

An example shows the accuracy gain that is available using iterative sinusoidal analysis instead of non-recursive analysis. The test signal consists of two sinusoids with frequencies 400 Hz and 600 Hz (sampling rate is 22050 Hz) with amplitudes 1 and 0.1, respectively. A hamming window with a length of 46.4 ms (1023 samples) is used. Figure 3 depicts the spectral representation of the windowed signal segment. The peaks are well-separated and the difference of magnitudes is 20 dB.

In this example we are interested in accuracy of parameter estimation on the sinusoid at

600 Hz when non-recursive and recursive sinusoidal analysis are used. In non-recursive modeling, the peaks are located directly at the magnitude spectrum of Figure 3, the location and magnitude estimates are fine-tuned using parabolic interpolation (see, e.g., Smith and Serra (1987)), and the phase value is linearly interpolated from the phase spectrum. In iterative modeling, the higher peak at 400 Hz is first detected, its parameters are estimated and a sinusoid is synthesized. The synthetic component is subtracted from the original signal and a first-order residual is obtained. The parameters of the weaker sinusoid are detected in the residual signal.

The same experiment was conducted using two similar test signals that had the sinusoid at 600 Hz with amplitudes of 0.01 and 0.5 corresponding to magnitude differences of 40 dB and 6 dB, respectively. Table 1 presents the results of the experiment. The results are sorted in rows according to the magnitude difference of the two sinusoids. The second and the third column tabulate the error in frequency estimation without and with iterative analysis, and the fourth and the fifth column show the amplitude estimation error without and with iterative analysis, respectively. The residual signals that are studied here are obtained by subtracting the synthesized weaker sinusoid from the original weaker sinusoid. The sixth and seventh column show the energies of the residual signals. Finally, the ratios of the residual energies are presented in the eighth column.

This example shows the improvement in accuracy that may be gained by using an iterative analysis. With 40 dB magnitude difference of the sinusoids, the improvement is most pronounced. Surprisingly, even when the magnitude difference is only 6 dB, the improvement is clear. Note that in some applications the accuracy gain obtained with iteration is not critical and the computationally more efficient non-iterative analysis may be preferred.

2.1 Significance Measure for Detected Sinusoids

In iterative analysis, recursion is carried on until no significant sinusoidal components are found in the current analysis frame. In audio coding applications where the main attempt is to minimize the signal bandwidth and where a psychoacoustic model is available, the significance criterion is straightforward: detect components until according to the psychoacoustic model no significant components exist. However, in applications that alter the signal in synthesis stage or synthesize only a part of the analyzed signal, e.g., in sound source separation, the psychoacoustic model simulating auditory masking may be too tight a criterion. For instance, a partial of a harmonic tone may be masked by partials of other harmonic tones in the mixture signal while when separated, the excluded partial may make an audible difference in the original and the synthesized signals.

Another simple method for estimating the significance of a partial is magnitude response thresholding. A preferably frequency-dependent threshold value is determined and all peaks that do not exceed the threshold are excluded. The main problem with this method is adaptive determination of the threshold. One can also observe the difference of energy of the residual signal in consecutive steps. When the difference is sufficiently small, i.e., when removing a sinusoid makes little difference in the residual, the iteration is stopped.

Determination of the significance criterion is closely related to the model order estimation problem in detection of parametric line spectra (Kay, 1988; Stoica and Moses, 1997). Methods for model order estimation include the use of the Akaike information criterion and autocorrelation matrix eigenvalues. Ali (1995) and Hamdy et al. (1996) apply a frequency estimation method based on works of Slepian (1978) and Thompson (1982). They have proposed another test for locating relevant sinusoidal components in the spectral representation (Ali, 1995).

We propose the following measure for significance of a sinusoidal component that is suited for iterative analysis of sinusoidal parameters:

$$R_k = \left| \frac{\mathbf{r}_k^T \mathbf{s}_k}{\mathbf{s}_k^T \mathbf{s}_k} \right| \tag{1}$$

where $\mathbf{r}_k$ is the residual signal after the kth sinusoid has been removed and $\mathbf{s}_k$ is the synthesized kth sinusoid. Equation 1 essentially measures the correlation between the residual signal and the synthesized sinusoid and scales it with the energy of the synthesized sinusoid. If the model of the sinusoid is accurate, we expect R_k to be small.

Some insight to the behavior of the significance measure may be gained by considering a simple example. The test signal consists of a single sinusoid with a known amplitude, frequency and phase:

$$x_{\text{test}}(n) = a\cos(\omega n + \phi), \quad n = 0, \ldots, N_{\text{win}} - 1.$$

We assume that the estimated parameters produce a synthesized signal

$$s_{\text{test}}(n) = (a + \Delta a)\cos(\omega(n + \Delta n) + \phi + \Delta\phi), \quad n = 0, \ldots, N_{\text{win}} - 1$$

where Δa, $\Delta\omega$, and $\Delta\phi$ are the errors in amplitude, frequency, and phase, respectively. The plots from the left to the right in Figure 5 show the value of the significant measure as a function of amplitude, frequency, and phase error, respectively. In each figures, the other two errors are assumed zero. Naturally, these plots only show the behavior of the significance measure along the axis of the three-dimensional space and they do not provide information when all the errors are non-zero.

More insight to the significance measure may be gained by considering how the measure is related to the energies of the original, sinusoidal, and residual signals. The original signal $\mathbf{x}$ may be expressed as a sum of the residual signal of the first iteration and the corresponding sinusoid, i.e.,

$$\mathbf{x} = \mathbf{r}_1 + \mathbf{s}_1.$$

The energy of the original signal is

$$\mathbf{x}^T\mathbf{x} = (\mathbf{r}_1 + \mathbf{s}_1)^T(\mathbf{r}_1 + \mathbf{s}_1) = \mathbf{r}_1^T\mathbf{r}_1 + \mathbf{s}_1^T\mathbf{s}_1 + 2\mathbf{r}_1^T\mathbf{s}_1.$$

The significance measure may thus be expressed as

$$R_1 = \left| \frac{\mathbf{r}_1^T \mathbf{s}_1}{\mathbf{s}_1^T \mathbf{s}_1} \right| = \left| \frac{\mathbf{x}^T\mathbf{x} - \mathbf{r}_1^T\mathbf{r}_1 - \mathbf{s}_1^T\mathbf{s}_1}{2\mathbf{s}_1^T\mathbf{s}_1} \right|$$

which shows that R_1 measures the energy difference of the original signal and the sinusoid and the residual signals relative to the energy of the sinusoidal signal. It is evident that when energies of the sinusoidal and the residual signal sum to that of the original signal the measure is zero.

2.2 Iterative Analysis Using Interpolated Parameter Trajectories

A typical problem with the iterative analysis model of Figure 2 is that sinusoids are synthesized with constant amplitudes and frequencies in the iteration although after the parameter trajectories have been formed, the instantaneous amplitudes and phases of the sinusoids are interpolated between the frames. Thus, the iterative analysis scheme using constant parameters does not fully exploit the initial assumption made for the sinusoids, namely, that the sinusoids are slowly varying in amplitude and frequency. In fact, the DFT provides parameters for the middle point of the analysis frames and they are extrapolated to obtain instantaneous amplitude and phase for that frame.

In (Edler et al., 1996), linearly varying frequency was permitted but the amplitudes of the sinusoids were required constant or forced to follow and amplitude envelope detected on the mixture signal. In the following, we present an analysis algorithm that uses interpolated parameter trajectories also in the iterative parameter detection. The proposed algorithm is motivated by more accurate parameter estimation due to decreased errors in the residual signal in each iteration. Furthermore, the algorithm together with the significance measure presented in the previous subsection may be used to decide on the accuracy of each synthesized sinusoid in each analysis frame.

In the proposed method, the previous and the next analysis frames are incorporated in iterative analysis of the parameters of the sinusoids in the current analysis frame. Figure 6 illustrates the principle. In each iteration, the most prominent component is first detected in the current frame. Its parameters are then detected in the previous and the next frame, and a sinusoid is synthesized using the parameters at three points, i.e., at the middle of the previous, current, and next frame. The time span of the synthesized sinusoid is shown with a dashed line. The only extrapolation of the parameter trajectories takes place at the end of the next frame.

The performance of the normal iterative analysis and the iterative analysis using interpolated parameter trajectories are compared in an example presented in Figure 7. Now the test signal is a sinusoid with linearly changing amplitude and frequency. The amplitude is changed from 1 to 0 and the frequency is changed from 400 to 540 Hz in 450 ms. The top plot presents a segment of the test signal. The residual signal of the normal iterative analysis is plotted in the middle, and that with the trajectory interpolation in the bottom. Notice the different amplitude scales in the figure. As expected, the trajectory interpolation significantly reduces the residual signal. In this case, the difference in the residual amplitude is approximately an order of magnitude.

The values of R for the iterative analysis without and with interpolated parameter trajectories are 0.0082, 0.0013, respectively. The ratio of the energies of the residual is 17.6

dB. This indicates that the analysis using interpolated parameter trajectories yields more accurate results than analysis using constant instantaneous frequency and amplitude envelope.

3 High-Resolution Estimation of Parameters of Colliding Sinusoids

Methods based on sinusoidal modeling have been proposed for separation of sound source contributions in musical duet signals (Maher, 1990) and for suppression of co-channel interference in speech signals (Quatieri and Danisewicz, 1990). Both of these studies report problems related to the frequency resolution of the discrete Fourier transform (DFT). The problem is particularly pronounced when two sinusoids corresponding to partials of two tones of different fundamental frequency are close to each other in frequency. In this case the sinusoidal components may not be resolved and the detected parameters are typically inaccurate.

An iterative method for estimation of sinusoidal parameters has been proposed by Depalle and Tromp (1996); Depalle and Hélie (1997). With that method, the analysis window size is reducible to 1.5 periods of the lowest frequency (frequency resolution).

In the following, we represent another technique that may be used for increasing frequency resolution. As explained in the next subsection, the technique is based on fitting a model of two sinusoids with closely spaced frequencies to the analysis signal.

3.1 Nonlinear Least-Squares Estimation of Colliding Sinusoids

The proposed technique in separation of colliding sinusoids is based on *nonlinear least-squares* (NLS) method in a relatively small vicinity in the frequency space. NLS is the most accurate (minimum-variance) unbiased method for estimating sinusoids in additive Gaussian white noise (Stoica and Moses, 1997; Kay, 1988). In the Gaussian noise case, it can be shown to equal the *maximum likelihood estimator* (MLE). Even when the noise process is not white, the estimator gives consistent estimates. It is thus better suited for the problem at hand compared to, e.g., other high-resolution frequency estimation methods, such as MUSIC and ESPRIT (Stoica and Moses, 1997; Kay, 1988). Furthermore, the general NLS estimator is easily modified to perform search only in the desired vicinity.

The basic idea is to apply the estimator locally in a vicinity that is determined from analysis of the fundamental frequencies. Global application of the estimator is infeasible since that would involve a highly nonlinear search over a high-dimensional parameter space. In this local application, the parameter space is essentially two-dimensional and the search space may be defined in advance for faster convergence. Furthermore, the estimates of the fundamental frequencies provide initial values for the search algorithm.

The global nonlinear least squares approximation has the following cost function (Stoica

and Moses, 1997)

$$G(f, a, \phi) = \sum_{n=0}^{N-1} \left| y(n) - \sum_{k=1}^{K} a_k e^{i(2\pi f_k n + \phi_k)} \right|^2 \tag{2}$$

where f is a vector of normalized frequencies of the sinusoids, a is a vector of amplitudes, ϕ is a vector of initial phases, N is the length of the analysis frame, and K is the number of sinusoids. The parameters that minimize function $G(f, a, \phi)$ determine the sinusoidal model that produces the optimal least-squares estimate of the observed data $y(n)$.

The cost function of the local model is derived from Equation 2 as

$$G(f, a, \phi) = \sum_{n=0}^{N-1} \left| y(n) - \sum_{k=1}^{2} a_k e^{i(2\pi f_k n + \phi_k)} \right|^2, \; f_1 \in [f_{1,\min}, f_{1,\max}], \; f_2 \in [f_{2,\min}, f_{2,\max}] \tag{3}$$

where the ranges $[f_{1,\min}, f_{1,\max}]$ and $[f_{2,\min}, f_{2,\max}]$ are pre-determined from the estimates of the corresponding fundamental frequencies.

As shown by, e.g., Stoica and Moses (1997), the cost function of Equation 3 is minimized with the following separated equations

$$\begin{aligned} \hat{f} &= \arg\max_f [Y^H B (B^H B)^{-1} B^H Y] \\ \hat{\beta} &= (B^H B)^{-1} B^H Y|_{f=\hat{f}} \end{aligned} \tag{4}$$

where

$$\begin{aligned} \beta_k &= a_k e^{i\phi_k} \\ \beta &= [\beta_1 \; \beta_2]^T \\ Y &= [y(0) \cdots y(N-1)]^T \\ B &= \begin{bmatrix} 1 & 1 \\ e^{i2\pi f_1} & e^{i2\pi f_1} \\ \vdots & \vdots \\ e^{i2\pi(N-1)f_1} & e^{i2\pi(N-1)f_1} \end{bmatrix} \end{aligned}$$

Given the initial approximations of the frequencies, an iterative optimization algorithm, such as the Newton method, may be used to maximize the first equation of (4) and thus obtain the frequencies of the sinusoids. After the frequencies are obtained, the amplitudes and initial phases are computed using the second equation of (4).

The examples in Figures 8–10 illustrate local NLS frequency estimation of two closely spaced sinusoids. The two sinusoids are of equal amplitude and have frequencies of 400 and 420 Hz. The analysis frame length is 46.4 ms giving $1/N$ frequency resolution of 21.5 Hz (the frame length is 1023 in samples at a sampling rate of 22050 Hz). In the example in Figure 8, the test signal consists only of the two sinusoids. The magnitude spectrum of the Hamming-windowed test signal is presented on the left. The Hamming-windowed DFT does not provide a sufficient frequency resolution for the two sinusoids to be separated, and a single broad peak is produced near 400 Hz. The plot on the right

shows a close-up of the magnitude spectrum. The two dashed vertical lines at 400 and 420 Hz present the actual frequencies of the two sinusoids. The two solid vertical lines almost coinciding with the dashed lines depict the frequency estimates obtained using the local NLS estimator at 400.2 and 419.8 Hz. In this case, the amplitude errors are 0.3 % and 0.1 % and the phase errors are 0.04 and 0.03 radians for the sinusoids at 400 and 420 Hz, respectively. The values of R for the first and the second sinusoid are $R_1 = 0.0034$, $R_2 = 0.0008$, respectively.

In the next example, a correlated noise signal is added to the test signal. The noise is produced by filtering white Gaussian noise by an allpole filter. The filter parameters are obtained with linear prediction from a Finnish vowel /a/ spoken by a male. The filter order is 20. Correlated noise was used since many parametric frequency estimator methods are dependent on the assumption that the additive noise is uncorrelated. However, as shown by the example of Figure 9 and suggested in, e.g., (Stoica and Moses, 1997; Kay, 1988), the NLS method does not rely on such an assumption. That is an attractive feature in audio applications where the noise components are rarely uncorrelated. The signal-to-noise ratio is 7.6 dB. The plot on the left again shows the magnitude spectrum of the signal up to 2000 Hz. As before, the two sinusoids result in a single broad peak. In addition, numerous spurious peaks appear in the magnitude spectrum as a result from the additive noise. The plot on the right presents a close-up to the two sinusoids. The solid lines are close to the dashed lines indicating a relatively accurate estimation of the frequencies of the sinusoids. The detected frequencies are 400.1 and 420.9 Hz. Now the amplitude errors are 2.7 % and 4.2 % and phase errors are 0.16 and 0.11 radians for the sinusoids at 400 and 420 Hz, respectively. It is evident that the noise increases estimation error but it does not prohibit the identification of two components. Now the values of the significance measure are $R_1 = 0.022$ $R_2 = 0.042$ for the first and the second sinusoid, respectively.

The third example of the local NLS parameter estimator is presented in Figure 10. In this experiment, a harmonic tone is added to the test signal of the previous example. The fundamental frequency of the tone is 265 Hz. The left plot shows the magnitude spectrum of the test signal. The the spectrum exhibits the peak corresponding to the two sinusoids near 400 Hz, peaks corresponding to the harmonic tones at integral multiples of the fundamental frequency, and spurious peaks due to the correlated noise. The close-up on the right indicates that the estimates of the frequencies of the sinusoids are again relatively accurate. The estimated frequencies are 400.2 and 418.8 Hz and the corresponding amplitude errors are 4.0 % and 0.5 % and phase errors are 0.15 and 0.17 radians, respectively. Again the NLS frequency estimator is able to distinguish the two sinusoids and provide reasonable estimates of their parameters. In this case, $R_1 = 0.033$ and $R_2 = 0.012$.

4 Sound Source Separation Example

The following example illustrates the use of sinusoidal modeling in sound source separation in a system of Figure 1. In this example, the pre-analysis block in Figure 1 is a multi-pitch analyzer that is reported in (Karjalainen and Tolonen, 1999). The multi-pitch

analysis model essentially divides the signal into two channels, below and above 1000 Hz, computes a "generalized" autocorrelation of the low-channel signal and of the envelope of the high-channel signal, and sums the autocorrelation functions. The summary autocorrelation function (SACF) is further processed to obtain an enhanced SACF (ESACF) which suppresses spurious periodicities. The SACF and ESACF representations are used in observing the (multiple) periodicities of the composite signal. The parameters of the model have been tuned experimentally to obtain good separation of typical harmonic complex mixtures.

The most prominent periodicities are detected in the ESACF representation and used in grouping of the sinusoidal components that have been detected using the iterative sinusoidal analysis algorithm using interpolated parameter trajectories. The sinusoidal components of the most prominent tone in each frame are subtracted from the original signal.

Figure 11 depicts the ESACF representation of each step of the iterative sound source separation. The test signal is a polyphonic excerpt of music by the classical guitar. Time is presented on the horizontal axis and periodicity lag on the vertical axis. The gray scale represents the prominence of periodicity. The ESACF of the test signal before any sound source have been separated is plotted on the top of Figure 11. A prominent melody of six notes around a lag of 2 ms (corresponding the a fundamental frequency of 500 Hz) may be identified in the ESACF representation. The trajectory of the fundamental frequency of this melody line has been detected and used in grouping of the sinusoidal components so that a representation of the harmonic tones of the melody line is obtained. The ESACF of the sinusoidal representation is depicted in the middle plot. The sinusoidal model is subtracted from the original signal and a residual signal is obtained. The ESACF of the residual signal is plotted on the bottom of Figure 11.

The plots of Figure 11 show the potential of iterative analysis. In the top plot, the prominent melody masks the other components in the ESACF representation. When the sinusoidal model is subtracted, the other tones become visible in the ESACF of the residual signal, as can be seen at the bottom of the figure. It is now possible to detect the most prominent tones in the residual signal and remove them in the next iteration.

5 Summary and Conclusions

We have discussed methods for separation of harmonic sound sources using sinusoidal modeling. The proposed techniques are 1) a local nonlinear least-squares (NLS) frequency estimator for sinusoids that are closely spaced in frequency, 2) an iterative analysis algorithm using interpolated parameter trajectories, and 3) a measure for testing significance and accuracy of detected sinusoids. An example was presented that shows how the methods may be used in sound source separation.

While the separation techniques perform reasonably well with sound signals that have prominent harmonic tones, they are not yet sufficiently robust to yield reliable separation

in the general case. In most cases, however, iterative sinusoidal analysis and recursive removal of prominent harmonic tones are able to detect and separate more harmonic tones with better accuracy than a one-run separation algorithm.

Acknowledgments

Vesa Välimäki is gratefully acknowledged for discussions and suggestions for the manuscript.

This work has been supported by the GETA Graduate School at Helsinki University of Technology, the Foundation of Jenny and Antti Wihuri (Jenny ja Antti Wihurin rahasto), and Nokia Research Center.

References

Ali, M. 1995. *Adaptive Signal Representation with Application in Audio Coding*, PhD thesis, University of Minnesota, Minneapolis, USA.

Anderson, D. V. 1996. Speech analysis and coding using a multi-resolution sinusoidal transform, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1037–1040.

Barrett, R. F. and Holdsworth, D. A. 1993. Frequency tracking using hidden Markov models with amplitude and phase information, *IEEE Transactions on Signal Processing* **41**(10): 2965–2976.

Depalle, P. and Hélie, T. 1997. Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows, *Proceedings of the IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Depalle, P. and Tromp, L. 1996. An improved additive analysis method using parametric modelling of the short-time Fourier transform, *Proceedings of the International Computer Music Conference*, Hong Kong, pp. 297–300.

Depalle, P., García, G. and Rodet, X. 1993. Tracking of partials for additive sound synthesis using hidden Markov models, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, Minneapolis, Minnesota, USA, pp. 225–228.

Ding, Y. and Qian, X. 1997. Processing of Musical Tones Using a Combined Quadratic Polynomial-Phase Sinusoid and Residual (QUASAR) Signal Model, *Journal of the Audio Engineering Society* **45**(7/8): 571–584.

Edler, B. and Purnhagen, H. 1998. Concepts for hybrid audio coding schemes based on parametric techniques, *Proceedings of the 104th Convention of the Audio Engineering Society*, Amsterdam, the Netherlands. Preprint 4808.

Edler, B., Purnhagen, H. and Ferekidis, C. 1996. ASAC—analysis/synthesis audio codec for very low bit rates, *Proceedings of the 102nd Convention of the Audio Engineering Society, Preprint 4376*, Copenhagen, Denmark.

George, E. B. and Smith, M. J. T. 1992. Analysis-by-synthesis overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones, *Journal of the Audio Engineering Society* **40**(6): 497–516.

George, E. B. and Smith, M. T. J. 1997. Speech analysis/synthesis and modification using an analysis-synthesis/overlap-add sinusoidal model, *IEEE Transactions on Speech and Audio Processing* **5**(5): 389–406.

Goodwin, M. 1997. *Adaptive Signal Models: Theory, Algorithms, and Applications*, PhD thesis, University of California, Berkeley.

Hamdy, K. N., Ali, M. and Tewfik, A. H. 1996. Low bit rate high quality audio coding with combined harmonic and wavelet representations, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1045–1048.

Karjalainen, M. and Tolonen, T. 1999. Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona. To appear.

Kay, S. M. 1988. *Modern Spectral Estimation: Theory and Application*, Prentice Hall, Englewood Cliffs, New Jersey, p. 543.

Levine, S. 1998. *Audio Representations for Data Compression and Compressed Domain Processing*, PhD thesis, Stanford University, CCRMA, Stanford, CA.

Levine, S. and Smith, J. O. 1998. A sines+transient+noise audio representation for data compression and time/pitch-scale modifications, *Proceedings of the 105th Convention of the Audio Engineering Society*, New York. Preprint 4781.

Levine, S. N., Verma, T. S. and Smith, J. O. 1997. Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio, *Proceedings of the IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Levine, S., Verma, T. and Smith, J. O. 1998. Multiresolution sinusoidal modeling for wideband audio with modifications, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 6, Seattle, USA, pp. 3573–3576.

Macon, M. W. 1996. *Speech Synthesis Based on Sinusoidal Modeling*, PhD thesis, Georgia Institute of Technology, Atlanta, Georgia, USA, p. 153.

Maher, R. C. 1990. Evaluation of a method for separating digitized duet signals, *Journal of the Audio Engineering Society* **38**(12): 956–979.

Masri, P. and Bateman, A. 1996. Improved modelling of attack transients in music analysis-resynthesis, *Proceedings of the International Computer Music Conference*, Hong Kong, pp. 100–103.

McAulay, R. J. and Quatieri, T. F. 1986. Speech analysis/synthesis based on a sinusoidal representation, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **34**(6): 744–754.

Prandoni, P., Goodwin, M. and Vetterli, M. 1997. Optimal time segmentation for signal modeling and coding, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, Muenchen, Germany, pp. 2029–2032.

Quatieri, T. F. and Danisewicz, R. G. 1990. An approach to co-channel talker interference suppression using a sinusoidal model for speech, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **38**(1): 56–69.

Rodríguez-Hernández, M. A. and Casajús-Quirós, F. J. 1994. Improving time-scale modification of audio signals using wavelets, *Proceedings of the 1994 International Conference on Signal Processing Applications & Technology*, Dallas, USA.

Serra, X. 1989. *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition*, PhD thesis, Stanford University, California, USA, p. 151.

Serra, X. and Smith, J. O. 1990. Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition, *Computer Music Journal* **14**(4): 12–24.

Slepian, D. 1978. Prolate spheroidal wave functions, Fourier analysis, and uncertainty – V: the discrete case, *Bell System Technical Journal* **57**(5): 1371–1340.

Smith, J. O. and Serra, X. 1987. PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation, *Proceedings of the International Computer Music Conference*, Urbana-Champaign, Illinois, USA, pp. 290–297.

Stoica, P. and Moses, R. 1997. *Introduction to Spectral Analysis*, Prentice Hall, Upper Saddle River, New Jersey, p. 319.

Streit, R. L. and Barrett, R. F. 1990. Frequency line tracking using hidden Markov models, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **38**(4): 586–598.

Thompson, D. J. 1982. Spectrum estimation and harmonic analysis, *Proceedings of the IEEE* **79**(9): 1055–1096.

Verma, T. S., Levine, S. N. and Meng, T. H. Y. 1997. Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals, *Proceedings of the International Computer Music Conference*, Thessaloniki, Greece, pp. 164–167.

Figures and Tables

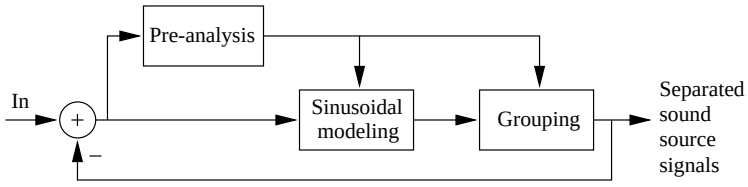

Figure 1: A block diagram of sound source separation system based on sinusoidal modeling.

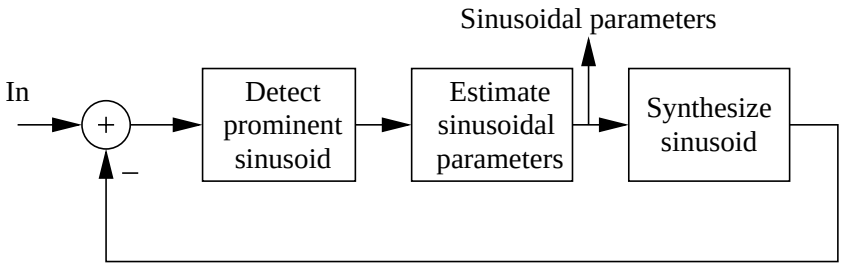

Figure 2: A block diagram of iterative sinusoidal analysis.

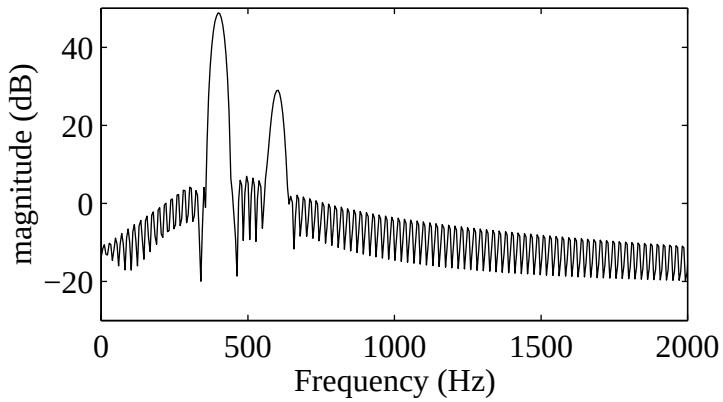

Figure 3: Spectrum of the test signal used in an example of iterative sinusoidal modeling. The test signal consists of two sinusoidal componensts with frequencies 400 and 600 Hz and amplitudes 1 and 0.1.

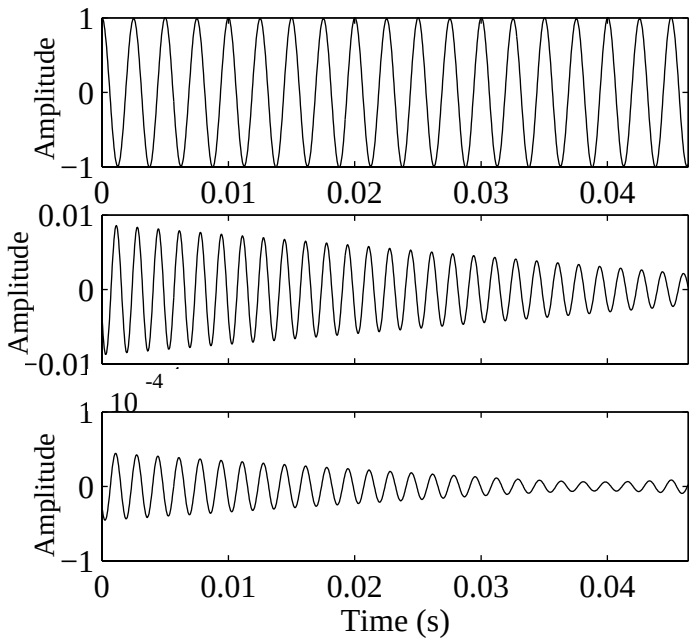

Figure 4: An example of iterative sinusoidal parameter estimation. Top: the weaker sinusoidal component of the test signal, middle: residual of the weaker sinusoid after regular sinusoidal modeling, and bottom: residual of the weaker sinusoid with iterative sinusoidal modeling.

	Amplitude		Frequency		Energy		
Magn. diff.	$\Delta f_{\mathrm{no-iter}}$	Δf_{iter}	$\Delta a_{\mathrm{no-iter}}$	Δa_{iter}	$E_{\mathrm{no-iter}}$	E_{iter}	$E_{\mathrm{iter}}/E_{\mathrm{noiter}}$
6 dB	0.1776	0.0076	0.0028	8.6546e-05	0.0390	1.2905e-04	0.0033
20 dB	0.7585	0.0022	0.0031	7.4245e-06	0.0352	4.1625e-06	1.1837e-04
40 dB	3.1197	0.0583	0.0040	1.0469e-05	0.0196	1.5689e-06	7.9867e-05

Table 1: Iteration results.

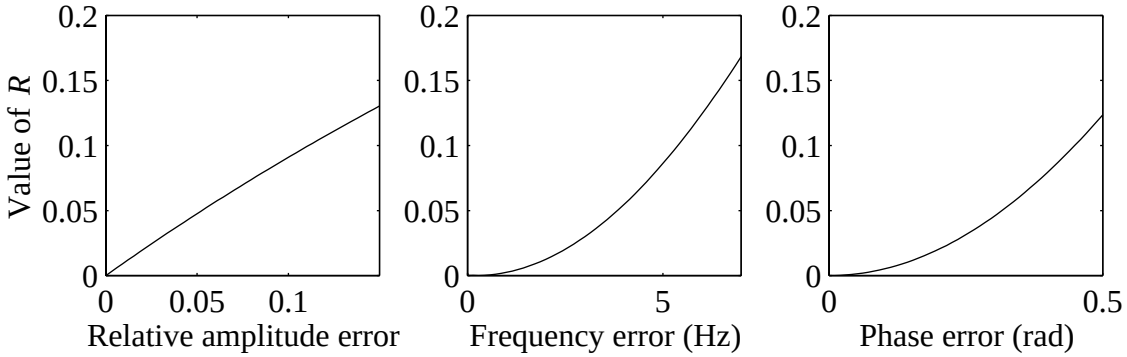

Figure 5: An example of the significance measure. The value of R is shown as a function of amplitude error ($\Delta\phi = 0$ and $\Delta\omega = 0$) in the left plot, as a function of frequency error ($\Delta a = 0$ and $\Delta\phi = 0$) in the middle plot, and as a function phase error ($\Delta a = 0$ and $\Delta\omega = 0$) in the right plot.

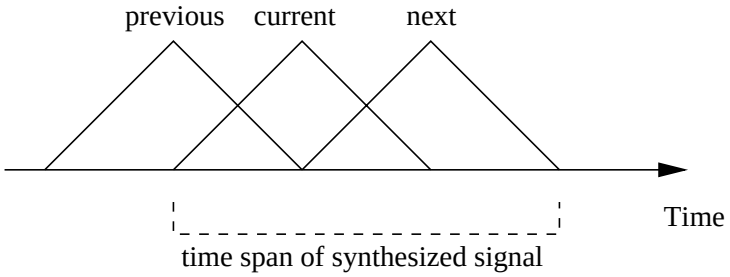

Figure 6: A schematic presenting the principle of iterative sinusoidal analysis with interpolated parameter trajectories.

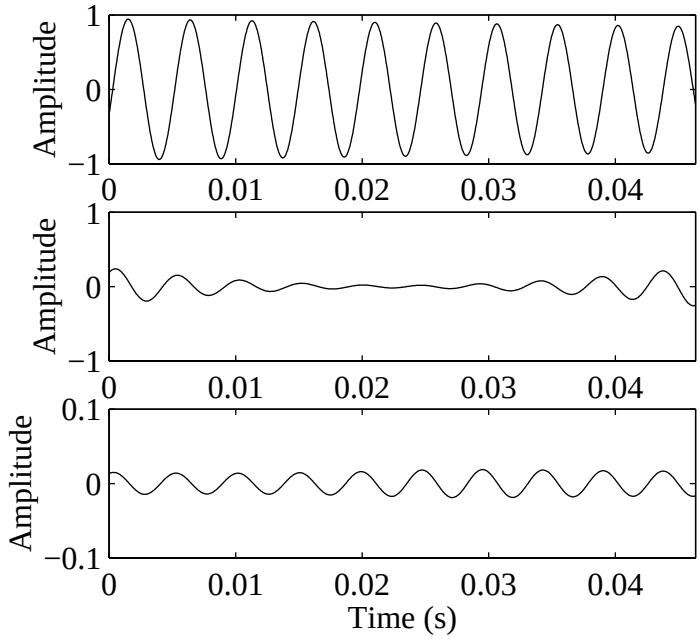

Figure 7: An example of interpolated iterative sinusoidal parameter estimation. Top: the test signal; middle the residual signal after iterative sinusoidal modeling using extrapolated parameter trajectories, and bottom: the residual using interpolated parameter trajectories.

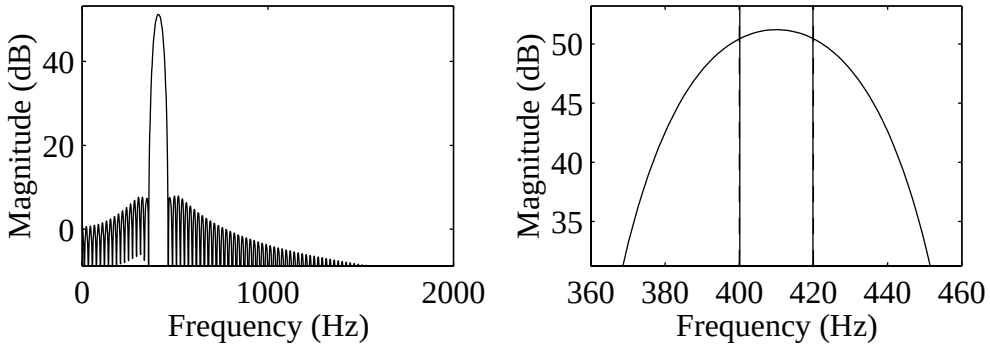

Figure 8: An example of the local nonlinear least-squares estimation of sinusoidal parameters. The test signal consists of two sinusoids with equal amplitudes and frequencies 400 and 420 Hz.

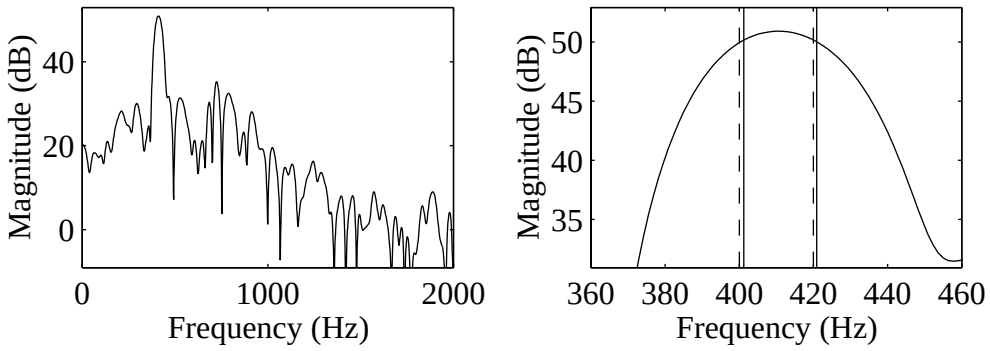

Figure 9: An example of the local nonlinear least-squares estimation of sinusoidal parameters. The test signal consists of two sinusoids with equal amplitudes and frequencies 400 and 420 Hz and correlated noise.

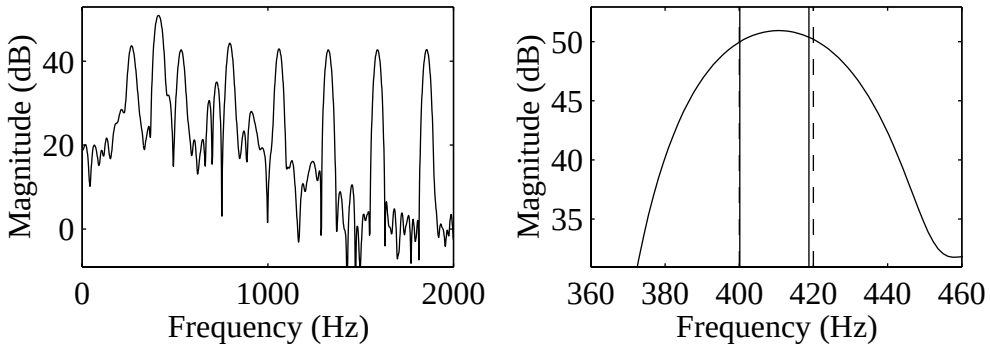

Figure 10: An example of the local nonlinear least-squares estimation of sinusoidal parameters. The test signal consists of two sinusoids with equal amplitudes and frequencies 400 and 420 Hz, additive correlated noise, and a harmonic tone.

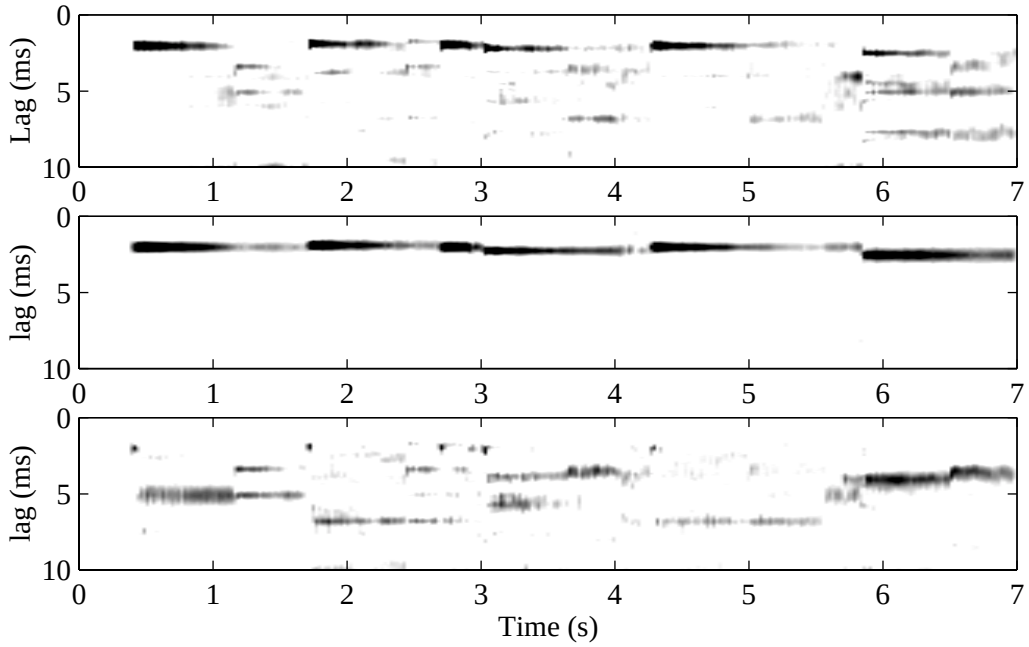

Figure 11: An example of separation of harmonic sound sources. The ESACF representation of the test signal is shown on the top. The ESACF of the sinusoidal model of a detected melody line is depicted in the middle. The bottom plot presents the ESACF of the residual signal.