CONSTRAINED TIME-VARIANT SIGNAL MODELING FOR IDENTIFYING COLLIDING HARMONICS IN SOUND MIXTURES

Miroslav Zivanovic[1] *and Johan Schoukens*[2]

[1]Dpt. IEE, Universidad Publica de Navarra, Campus Arrosadia, 31006, Pamplona, Spain
phone: + 34 948 16 90 24, fax: + 34 948 16 97 20, email: miro@unavarra.es

[1]Dpt. ELEC, Vrije Universiteit Brussel, Pleinlaan 2, B-1050, Brussels, Belgium
phone: +32 2 629 29 47, fax: +32 2 629 28 50, email: jschouk@vub.ac.be

ABSTRACT

We present a general approach to identifying time-variant colliding harmonics in pitched steady-state monophonic sound mixtures. Each sound is described by a linear-in-parameter quasi-harmonic model which captures properly instantaneous time variations of the non-stochastic sound energy. The model parameters corresponding to colliding harmonics are estimated on the basis of multiple-solution cost function L^2 minimization with regularization. The major advantage against the state-of-the art methods is that no additional information about the underlying sounds in the mixture is needed. A comparative study shows that the proposed method performs significantly better than an existing algorithm for separation of monaural pitched sounds from a mixture.

1. INTRODUCTION

Single-channel separation of monophonic harmonic sound sources from a polyphonic mixture is proven to be a useful analysis tool for characterizing complex audio signals like music and co-channel speech. Some of the most prominent application areas include information retrieval, polyphonic music transcription and structured audio coding.

In the past decade a number of approaches, which can generally be referred to as supervised and unsupervised, have been investigated. The supervised methods e.g. [1-3] are typically rooted in training individual source models on the basis of a large number of recordings. The unsupervised methods are more general and usually perform the separation by combining computational auditory scene analysis cues with either data-driven redundancy-reduction [4, 5] or parametric model-based inference [6-8]. The unsupervised methods often rely on the quasi-harmonic structure assumption for the underlying sound sources, provided that the corresponding fundamental frequencies (pitches) can be correctly estimated. In the best-case-scenario i.e. no frequency-domain overlap among the sources, those methods yield very good results with a minimal computational effort.

In real-world polyphonic music signals, however, the pitches of the mixed sources often form consonant intervals, which correspond to frequency relationship given as small integer ratios, namely, an octave (2:1), a fifth (3:2), a forth (4:3) and a third (5:4). As a result, a large number of harmonics gets clustered in a very small frequency bandwidth, thus reducing significantly the identifiability of the individual sources. Considerable efforts were made to address the problem of colliding harmonics in different ways, which include spectral filtering [6], time-frequency smoothness constraints [7] and common amplitude modulation from human auditory perception [8]. The common factor shared by these approaches is the use of single source information estimated from non-colliding harmonics and prior assumptions about sound harmonic structure. While the identification of the non-colliding harmonics is relatively easy task to perform, the prior assumptions of quasi-stationarity [6], spectral smoothness [7] or phase-locked harmonics [8] are often not justified for retrieving colliding harmonics in real-world music signals.

The approach we present herein combines single-source time-variant harmonic modeling and regularized least-squares for model parameter estimation. It can be seen as a specific extension of the harmonic model [9] oriented to sound source separation with an arbitrary number of colliding harmonics. Amplitude and frequency time variations of each source are efficiently captured by a single continuous polynomial which modulates a set of harmonically related sines and cosines. The model parameters for each source are estimated by minimizing a multi-objective L^2 cost function consisting of the reconstruction error term (all harmonics) and collision term (colliding harmonics). As the model is linear-in-parameters, the minimizer is a regularized least-squares (approximate) solution of the reconstruction term.

The present method brings out novelties in comparison to [6-8], namely: 1) simultaneous modeling of all the harmonics in the mixture, 2) time variations in all the harmonics are properly captured and 3) no information from the non-colliding harmonics are needed.

2. THE SIGNAL MODEL

A discrete steady-state sound mixture signal can be efficiently represented as a superposition of a number of monophonic harmonic sources $h_r(n)$, r = 1, 2, …, R and additive noise $e(n)$:

$$s(n) = h(n) + e(n) = \sum_{r=1}^{R} h_r(n) + e(n). \qquad (1)$$

Each source is typically modeled through a set of non-stationary harmonics (partials) while the noise is represented as a Gaussian process with time-varying spectral shape.

2.1. The time-variant harmonic model

Among a number of monophonic harmonic signal models that have been proposed in the recent years, we choose the one described in [9]. It has been proven to provide a compact description of amplitude and frequency time-variations in the analysis window by means of just a small number of parameters. Starting from the expression of a single quasi-harmonic source $h_r(n)$ of I components:

$$h_r(n)=\sum_{i=1}^{I} A_r^{(i)}(n)\cos\left[2\pi\, i f_{0r}(n)n+\varphi_r^{(i)}\right]=$$

$$=\sum_{i=1}^{I} a_r^{(i)}(n)\sin\left(2\pi\, i f_{0r}(n)n\right)+b_r^{(i)}(n)\cos\left(2\pi\, i f_{0r}(n)n\right),$$

$$a_r^{(i)}(n)=-A_r^{(i)}(n)\sin\varphi_r^{(i)},\quad b_r^{(i)}(n)=A_r^{(i)}(n)\cos\varphi_r^{(i)},\ (2)$$

the corresponding time-variant harmonic model is derived by using the assumptions of continuity and slow-variation:

1) *Continuity assumption*: $a_r^{(i)}, b_r^{(i)}, f_{0r}$ are modeled as a continuous order-P time polynomials:

$$\hat{a}_r^{(i)}(n)=\sum_{k=0}^{P} a_{rk}^{(i)} n^k\,,\qquad \hat{b}_r^{(i)}(n)=\sum_{k=0}^{P} b_{rk}^{(i)} n^k\,,$$

$$\hat{f}_{0r}(n)=\sum_{k=0}^{P}\frac{f_{kr}}{k+1}n^k\,. \qquad (3)$$

2) *Slow-variation assumption*: the following approximations are applicable:

$$\sin 2\pi\, i\frac{f_{kr}}{k+1}n^{k+1}\approx 2\pi\, i\frac{f_{kr}}{k+1}n^{k+1},$$

$$\cos 2\pi\, i\frac{f_{kr}}{k+1}n^{k+1}\approx 1\,. \qquad (4)$$

By combining (1) - (4) and after some mathematics as shown in [9], we obtain the time-variant harmonic model for a steady-state sound mixture of R sources:

$$\hat{h}(n)=\sum_{r=1}^{R}\sum_{i=1}^{I}\alpha_r^{(i)}(n)\sin\left(2\pi\, i f_{0r} n\right)+\beta_r^{(i)}(n)\cos\left(2\pi\, i f_{0r} n\right).\ (5)$$

The polynomials $\alpha_r^{(i)}$ and $\beta_r^{(i)}$, whose internal structure is herein omitted for the sake of clarity, capture the non-stationarities in RI harmonics of the model (5). The relationship among harmonics belonging to different sources is one of the key issues for sound source separation and will be tackled in the following sections.

3. THE SIGNAL MODEL ESTIMATION

In the present section we discuss the estimation problem in relation to harmonic classification and propose a method to successfully separate colliding and non-colliding harmonics from a harmonic mixture.

3.1. The linear system

The model (5) evaluated at N discrete time points can be represented as a system of linear equations of the following matrix form:

$$\boldsymbol{s}=\boldsymbol{\Phi\theta}+\boldsymbol{\varepsilon}\,. \qquad (6)$$

The vector $\boldsymbol{\varepsilon}$ comprises both modeling and measurement errors, the vector $\mathbf{s}$ represents the measurement data, the vector $\boldsymbol{\theta}$ contains $2RI(P+1)$ model parameters:

$$\boldsymbol{\theta}=\left(\boldsymbol{\theta}_1^{(1)}\ldots\boldsymbol{\theta}_1^{(I)}\boldsymbol{\theta}_2^{(1)}\ldots\boldsymbol{\theta}_2^{(I)}\ldots\boldsymbol{\theta}_R^{(1)}\ldots\boldsymbol{\theta}_R^{(I)}\right)^{\mathrm{T}},$$

$$\boldsymbol{\theta}_r^{(i)}=\left(\alpha_{r0}^{(i)}\ \alpha_{r1}^{(i)}\ldots\alpha_{rP}^{(i)}\ \beta_{r0}^{(i)}\ \beta_{r1}^{(i)}\ldots\beta_{rP}^{(i)}\right)^{\mathrm{T}}, \qquad (7)$$

and the N-entry matrix $\boldsymbol{\Phi}$ provides the model description:

$$\boldsymbol{\Phi}=\left(\phi_{1\alpha}^{(1)}\ldots\phi_{1\alpha}^{(I)}\ldots\phi_{R\alpha}^{(1)}\ldots\phi_{R\alpha}^{(I)}\ \phi_{1\beta}^{(1)}\ldots\phi_{1\beta}^{(I)}\ldots\phi_{R\beta}^{(1)}\ldots\phi_{R\beta}^{(I)}\right)^{\mathrm{T}},$$

$$\left(\phi_{r\alpha}^{(i)}\right)_{nm}=n^m\sin\left(2\pi i f_{0r}n\right),$$

$$\left(\phi_{r\beta}^{(i)}\right)_{nm}=n^m\cos\left(2\pi i f_{0r}n\right), \qquad (8)$$

where m = 1, 2, …, $2RI(P+1)$. For the sake of the present application, it is assumed that the mean fundamental frequencies f_{0r}, r = 1, 2, …, R are estimated before hand by means of a multipitch estimation algorithm like [10, 11]. Further discussion on multipitch estimation, however, is omitted as it is out of scope of the paper.

3.2. The ill-posedness of the system

In audio signal modeling the number of model parameters is generally smaller than the number of measurement points in the analysis window. Accordingly, the system (6) is inconsistent and with no exact solution. However, if the system has full-column rank, an approximate solution can always be found by computing the residual that minimizes some norm (typically L^2 norm):

$$\min_{\theta}\left\|\mathbf{s}-\boldsymbol{\Phi\theta}\right\|_2^2,\ \text{ with }\ \hat{\boldsymbol{\theta}}=\boldsymbol{\Phi}^{+}\mathbf{s}\,, \qquad (9)$$

where $\boldsymbol{\Phi}^{+}$ is the Moore-Penrose pseudoinverse of $\boldsymbol{\Phi}$. If there is no significant frequency-domain overlap between the underlying sources (only non-colliding harmonics), linear least-squares (LS) can provide a stable solution $\hat{\boldsymbol{\theta}}^{\mathrm{LS}}$ to the system, as it does for a monophonic single-source scenario. As shown in [9] the resulting approximation error ε in (6) is close to the measurement error e in (1), thus rendering the modeling errors almost negligible. Such system is called well-posed system.

In presence of colliding harmonics the system (6) undergoes certain modifications which attain its well-posedness. The major change concerns the model matrix **Φ,** where each column is a sine/cosine multiplied by a corresponding polynomial term (8). If the input includes colliding harmonics as well, the corresponding columns in **Φ** become highly correlated, thus yielding the matrix condition number (the largest singular value to the smallest one) extremely large. In addition, the singular values of the matrix gradually decay to zero, which means that there is no obvious way to filter out the smallest singular values and turn the problem into a nearly-consistent rank-deficient one [12]. If the input signal is noise-free, it would still be possible to achieve a physically meaningful solution in spite of the ill-posedness of the problem. In real-world signals, however, the stochastic component acts as a perturbation to the ill-posed system, providing highly instable LS solution which is usually far from the unperturbed LS solution.

3.3. The cost function and regularized LS

In order to keep the problem less sensitive to perturbations, we have to mitigate the ill-conditioning of the model matrix **Φ**. This is usually done by adding a regularization term proportional to the desired solution norm in the system minimization cost function. As we are dealing with linear systems (6), a natural choice is L^2-norm regularization (also known as Tikhonov regularization) which computes a solution of the following cost function:

$$\min_{\theta}\left\{\left\|\mathbf{s}-\mathbf{\Phi\theta}\right\|_2^2+\lambda^2\left\|\mathbf{L\theta}\right\|_2^2\right\},\ \lambda\geq 0,\ \lambda\in\Re. \quad (10)$$

The first term in the cost function is known as the residual norm while the second term is referred to as the solution norm. It is an extension of (9) with the objective to keep the system solution within a certain range of values.

Minimizing (10) with respect to **θ** we obtain the regularized LS solution:

$$\hat{\boldsymbol{\theta}}^{\mathrm{RLS}}=\mathbf{\Phi}_{\lambda}^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{s}=\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}+\lambda^2\mathbf{L}^{\mathrm{T}}\mathbf{L}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{s}. \quad (11)$$

The key issue in (11) is the presence of the additional term $\lambda\mathbf{L}^{\mathrm{T}}\mathbf{L}$ in the matrix $\mathbf{\Phi}_{\lambda}$. With an appropriate choice of the matrix **L** and scalar λ the conditioning of the matrix $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ can be substantially improved. Accordingly, this kind of regularization alleviates the perturbations in the solution and preserves desired features of the estimated solution.

3.4. The choice of regularization parameters

The parameters **L** and λ have each a particular roll in regularization of the solution norm. The matrix **L** should reflect an a priori knowledge (if available) about the solution. For example, if some information about the smoothness of the true solution is known, then **L** is typically constructed as a discrete derivative operator. In case that nothing is known beforehand about the true solution, **L** is often chosen as a unitary matrix.

From the previous subsections it is clear that ill-posedness of the linear system comes from the colliding harmonics. Therefore, the regularization in (10) should concern only the colliding harmonics in the mixture. Once the initial colliding/non-colliding classification has been carried out, we define **L** as follows:

$$\mathbf{L}=\mathrm{diag}\left(\mathbf{M}\circ\boldsymbol{\theta}\right),$$

where "∘" represents the Hadamard entrywise vector product and **M** is a binary masking vector defined as:

$$\mathbf{M}=\left(\mathbf{M}_1\ \mathbf{M}_2\ \ldots\mathbf{M}_{RI}\right)^{\mathrm{T}},$$

$$\mathbf{M}_i=\begin{cases}\left(\mathbf{1}\right)_{1\times 2(P+1)},\ i^{\mathrm{th}}\text{ harmonic is colliding}\\ \left(\mathbf{0}\right)_{1\times 2(P+1)},\ i^{th}\text{ harmonic is non - colliding}\end{cases}$$

Such choice of matrix **L** allows for simultaneous estimating of all harmonics in the mixture. Furthermore, it operates exclusively on the smallest singular values of **Φ**, thus guaranteeing well-posedness of $\mathbf{\Phi}_{\lambda}$.

The parameter λ controls the degree of regularization and is highly application-dependent. Too large values of λ yield an over-determined solution (the solution is strongly distorted) while λ too small does not achieve to control the perturbation sensitivity (under-determined solution). The determination of the optimal value of λ is a very difficult task, as it depends on the properties of **θ**, **Φ** and the perturbation variance. Some methods like L-curve, Generalized Cross-Validation (GCV) and the Discrepancy principle tend to produce undersmoothed estimates (especially in high-noise environment) and are usually time-consuming in the sense of computation effort.

In the context of the present method, we seek a simple approach to choosing an approximate λ that would keep the system computational simplicity and at the same time provide a close-to-optimal solution. We have carried out an empirical study regarding two colliding harmonics, where we have calculated the expected value of the norm error between the correct model parameters **θ** and its regularized estimate $\boldsymbol{\theta}^{\mathrm{RLS}}(\lambda)$ as a function of both harmonic frequency separation and λ. For a fixed Gaussian perturbation with standard deviation 0.1 (SNR = 15 dB) and harmonic separation expressed in terms of the analysis resolution $1/N$ (N is the size of the analysis window) the 2D error norm is plotted in Figure 1. The intensive bright area to the left in the figure (for λ approximately up to 0.2) corresponds to a large norm which is due to a very strong estimate variance. To the right, the norm increases more smoothly because it is dominated by the estimate bias. The dark area in the upper-left half represents the values of the norm which are situated in the neighborhood of the minimum. Such error norm constellation is preserved for the SNR as small as 5 dB approximately; further decrease in SNR implies a sharper error norm minimum and consequently a more selective choice of λ. Nevertheless, except for extremely small harmonic separation below $0.1/N$ and large

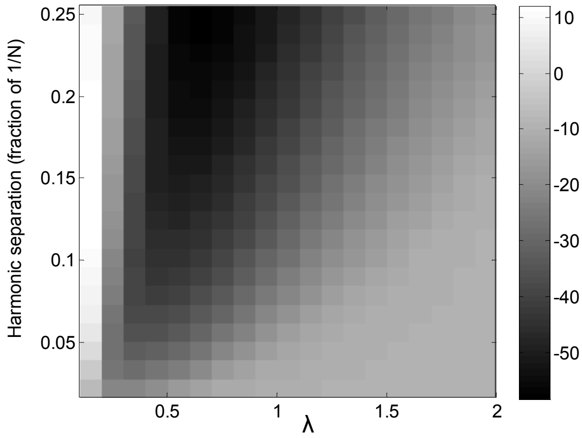

Figure 1 – Expected value of the error norm in decibels as a function of λ and harmonic separation (given as a fraction of $1/N$).

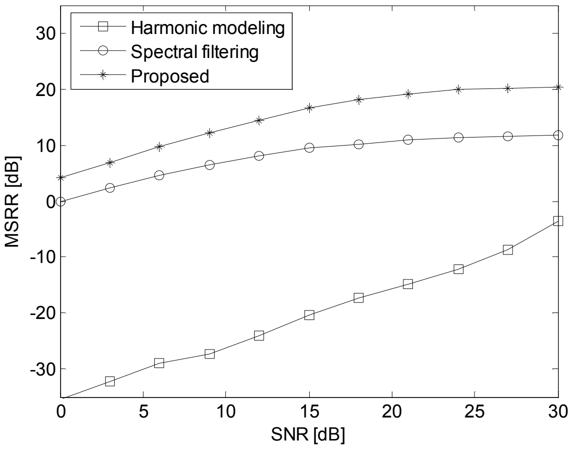

Figure 2 – Mean Signal-to-residual ratio for the methods.

noise variance (SNR < 5 dB), we can always choose $\lambda \in [0.5, 0.7]$.

However, the closeness of the estimated model parameters (7) to the correct ones with such λ must always be interpreted in the context of the desired application. If a user wants to characterize a sound source by a further analysis of the estimated parameters, then an additional research on sensitivity of sound properties against the error norm must be carried out. However, if the goal of the regularization is to separate the source waveforms from a mixture, then this choice of λ it will achieve a very good approximation quality, as discussed in the following section.

4. EXPERIMENTAL RESULTS

A comparative study, followed by an illustrative real-world example, is presented in order to discuss the performance of the proposed constrained time-variant harmonic modeling. Throughout this section we have used λ = 0.6.

For the comparative study we use as reference methods the Spectral filtering method [6] and the Harmonic modeling method [9]. The Spectral filtering method estimates a source from a mixture by analyzing the harmonic spectral peaks in the signal's STFT. For each set of colliding peaks a small-bandwidth spectral filter (type B) [6] is applied in order to split the signal's energy into its constituents. The Harmonic modeling method, which models each harmonic from the mixture independently of the rest according to Section 2.1, was not specifically designed to deal with colliding harmonics; nevertheless, it provides a clear picture of the signal model performance with and without regularization.

The test signal consists of two harmonic sources whose instantaneous frequencies and amplitudes vary according to sinusoidal quasi-vibrato laws plus Gaussian noise $r(n)$:

$$s_t(n) = \sum_{r=1}^{2}\sum_{i=1}^{15} A_r^{(i)}(n)\cos\left[2\pi\, i f_{0r} n + \varphi_r^{(i)}(n)\right] + r(n)$$

$$\varphi_r^{(i)}(n) = iA_{FM}\,\sin(2\pi\, f_{FM} n) + \gamma_r\,,$$

$$A_r^{(i)}(n) = (A_0/i)\left[1 + A_{AM}\,\cos(2\pi\, f_{AM} n + \delta_r)\right].$$

The fundamental frequencies f_{01} = 440 Hz and f_{02} = 660 Hz have been chosen in such a way to emulate the perfect fifth frequency interval, which in turn yields almost 40% of all harmonics to be collided. In addition, γ_r and δ_r assure arbitrary shape spectral envelopes for the underlying sources, in order to get as closer as possible to the real-world scenarios. The modulation denormalized frequencies are $f_{\mathrm{AM}}f_s$ = 22 Hz and $f_{\mathrm{FM}}f_s$ = 6 Hz, while the modulation amplitudes are $A_{AM} = A_{FM} = 0.2$. The analysis window length is 20 ms with the sampling frequency f_s = 44.1 kHz.

In order to measure the performance of the methods in a noisy environment, we vary the contribution of $r(n)$ through the SNR $\in$ [0, 30] dB in steps of 3 dB and for each SNR we calculate the Mean Signal-to-Residual Ratio (MSRR) over 500 realizations:

$$\mathrm{MSSR}_{dB} = 10\log_{10}\left[\sum_{r=1}^{2}\sqrt{\sum_{n=0}^{N-1} s_{tr}^2(n) \Big/ \sum_{n=0}^{N-1}\left(s_{tr}(n) - \hat{s}_{tr}(n)\right)^2}\right]$$

where $s_{tr}(n)$ and $\hat{s}_{tr}(n)$ are the original and approximated sources respectively. The resulting curves shown in Figure 2 follow a general ascending trend according to the increments in SNR. The proposed method clearly outperforms the reference methods in the whole analysis range. The Harmonic modeling is the worst due to the ill-posedness of the model in the neighborhood of the colliding harmonics. The Spectral filtering method achieves much better results because it processes the colliding harmonics in a non-parametric way. However, it remains around 5-8 dB below the proposed method. This is due to the initial assumptions about quasi-stationarity and spectral continuity of the underlying sources.

We illustrate the behavior of the proposed method by processing a mixture of two real-world monophonic sources (flute and trumpet) taken from the Iowa University music database [13]. While the flute features a quasi-stationary signal at 440 Hz, the trumpet performs a vibrato technique around 660 Hz (perfect fifth). In Figure 3 we observe a strong resemblance between the original and approximation spectrograms (plotted according to the same gray scale) except for a slight difference around 0.25 normalized time axis.

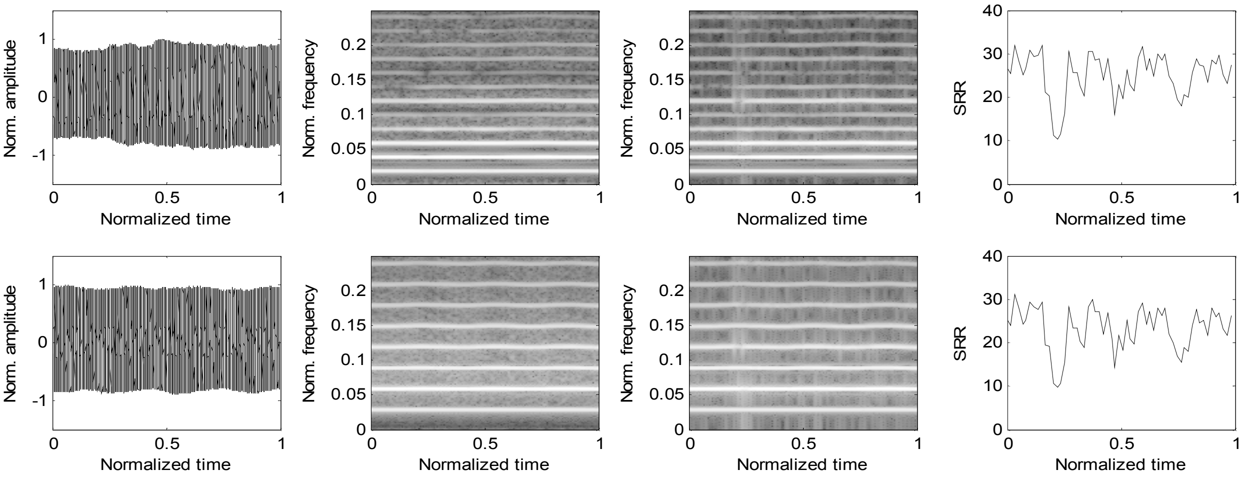

Figure 3 – First row: flute signal, its spectrogram, approximated spectrogram and signal-to-residual ratio; Second row: trumpet signal, its spectrogram, approximated spectrogram and signal-to-residual ratio.

A close examination of the harmonic frequencies for this specific time instant reveals that for three consecutive analysis frames the colliding harmonics are separated less than 1 Hz i.e. 0.02/*N*. According to Figure 1 we expect a certain increment in error norm and this is indeed reflected in the SRR reduction to 10 dB in Figure 3. Elsewhere, the SRR varies between 15-30 dB thus providing a very good source separation. Informal listening to a separated source reveals no audible residual from the other source and vice versa.

5. CONCLUSIONS

We have shown that constrained time-variant harmonic modeling is an appropriate tool for separating colliding harmonics in the context of monaural pitched sounds separation. With a simple choice of regularization parameters the proposed method can successfully identify colliding harmonics separated more than 5 % of the analysis resolution with SRR of about 20 dB. In addition, it outperforms the Spectral filtering method by 5-8 dB depending on the SNR. Furthermore, the system model is linear-in-parameters, which assures high computational efficiency and implementation simplicity.

REFERENCES

[1] M. Bay, J. W. Beauchamp, “Harmonic source separation using prestored spectra,” *Proceedings of the 6th ICA Workshop*, Charleston, USA, March 2006.

[2] E. Vincent, “Musical source separation using time-frequency source priors,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 1, pp. 91–98, January 2006.

[3] A. Ozerov, O. Philippe, R. Gribonval, F. Bimbot, “One microphone singing voice separation using source-adapted models,” *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* New Paltz, USA, 2005.

[4] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[5] R. Badeau, V. Emiya, B. Davis, “Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra**,**” *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 2009.

[6] M. R. Every, J. E. Szymanski, “Separation of synchronous pitched notes by spectral filtering of harmonics,” *IEEE Transactions on Audio, Speech and Signal Processing.* Vol. 14, No. 5, pp. 1845-1856, September 2006.

[7] T. Virtanen, A. Klapuri, “Separation of harmonic sounds using linear models for the overtone series,” *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, USA, 2002.

[8] J. Woodruff, Y. Li, D. Wang, “Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation,” *Proceedings of the International Conference on Music Information Retreival,* Phyladelphia, USA, September 2008.

[9] M. Zivanovic, J. Schoukens, “On the polynomial approximation for time-variant harmonic signal modeling,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 3, pp. 458-467, October 2010.

[10] A.P. Klapuri, “A perceptually motivated multiple-f0 estimation method for polyphonic music analysis,” *IEEE workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, October 2005.

[11] C. Yeh, A. Robel, X. Rodet, “Multiple fundamental frequency estimation of polyphonic music signals,” *IEEE, International. Conference on Acoustics, Speech and Signal Processing,* Philadelphia, USA, March 2005.

[12] W. Demmel, “Applied Numerical Linear Algebra,” SIAM, Philadelphia, 1997.

[13] http://theremin.music.uiowa.edu/MIS.html.