Optimal Filtering of an Instrument Sound in a Mixed Recording Using Harmonic Model and Score Alignment

Adiel Ben Shalom, Shai Shalev-Shwartz, Michael Werman and Shlomo Dubnov
School of Engineering and Computer Science, Hebrew University
chopin@cs.huji.ac.il
Center for Research in Computing and the Arts
sdubnov@ucsd.edu

Abstract

In this paper we show a method for optimal filtering (in terms of mean square error) of an instrument in a mixed recording. The filter is based on prior knowledge of the exact pitch played by the instrument and also on the assumption that the instrument is well described within the harmonic model framework. To get the pitch priors, we use alignment of the score information to the real recording. We show how this algorithm can be used to filter a single instrument or voiced singer. We also show how it can be used to substruct an instrument or to balance the volume to several instruments.

1 Introduction

This paper describes a method for filtering an instrument out of a mixed recording. This problem is very interesting and it has many applications. For example, in a multi-track recording one could re-mix the sound from the stereo recording without having the original tracks. Another future applications could be automatic karaoke and automatic music minus one.

Without any prior knowledge, this problem is considered very hard. An algorithm that would filter a single instrument out of a mixed recording, should be able to analyze precisely a complex musical scene. A lot of difficult problems should be solved such as polyphonic pitch detection, instrument recognition and more (Cook and J 1994; Virtanen and Klapuri 2002).

We solve this difficulty by using score alignment algorithm (Shalev-Shwartz, Dubnov, Friedman, and Singer 2002). We assume that the score is given along with the real recording. This is not a strong assumption since almost all popular and classical music can be found in MIDI format.The score alignment algorithm takes as input the real recording and the score information (MIDI) and output the precise alignment over time between them. This method simplifies the analysis of the complex sound.

Once an alignment is achieved, we use the exact time-pitch information to design an optimal filter for the different notes played by the different instruments. The filter that we describe is based on the harmonic model, and it is assumed that the instrument that we wish to filter can be modeled with the harmonic model. We explain in the next section the details of the filtering algorithm. We also show how this algorithm can be used to filter vocal voice (2.3), substruct an instrument out of a mixed recording (2.4) and also how it can be used to balance the volume between several instruments (2.5).

There are several drawbacks for this algorithm. First, it is limited for instruments that are well modeled within the harmonic model framework. It can not process percussion like instruments. We also need to have special processing for the attaca part of the sound as well as for unvoiced parts of a singer. Another drawback is that it cannot do source separation of unison(i.e if two instruments play the same pitch then algorithm cannot do any kind of filtering/separation between them)

2 Harmonic Filter Theory

2.1 Harmonic Model and Minimum Variance Estimation

The optimal filtering process is based on two concepts. The first is the *harmonic model* (Serra 1989; Rodet 1997; Roads, Pope, and Piccially 1997) assumption for the instrument pitch. The second is representing the filtering problem as a *Linear Constraint Minimum Variance* (LCMV) estimation problem. The LCMV approach is very natural in this setup, since the mixing instructions can be viewed as specification of a linear constraint on the resulting signal. In the LMCV approach, a filter design is based on assumption of a particular shape of the signal of interest (usually a single sinusoid). In our case, we assume that the signal of interest is described by a harmonic model.

Harmonic model means that the signal contains a periodic

pitch component comprising of a fundamental along with its partials, plus some noise that we model as white gaussian noise (WGN). Formally,

$$x(n) = \sum_{k=1}^{K} A_k \cos(2\pi k f_0 n + \phi_k(n)) + w(n) \qquad (1)$$

where A_k and ϕ_k are the instantaneous amplitude and phase of the k'th sinusoidal component K is the number of partials and $n = 0 \ldots, N-1$ where N is the number of samples.

The second concept that we adopt is Linear Constraint Minimum Variance estimation. Minimum variance signal estimation is a common method in classical parameter estimation theory (Kay 1993; Therrien 1992). The minimum variance parameter estimation depicts a method for estimating an unknown parameter in a noisy environment. The method guarantees to be optimal in the sense of minimizing the *mean square error* (MSE) for a given parametric model. If we denote the true value of the parameter to be estimated as θ and our estimator for the parameter as $\hat{\theta}$, then the MSE is defined as

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \qquad (2)$$

where E denotes expectation. It can be shown that the MSE is minimized for an unbiased estimator (MVU), which has the minimum variance of all possible estimators. Unfortunately, in many situations it is hard to find the MVU because it may require a prior knowledge about the distribution of the source. However, if it is possible to represent the estimation problem in a linear form, then we can take advantage of some unique properties of the linear data model and quite easily find the best minimum variance linear estimator.

2.2 Filtering a Single Instrument

We begin with the first scenario, where we wish to extract a single instrument from a musical recording whose score information is given. From the alignment process we get the segmentation information of this instrument in the recording which includes the exact pitch that the instrument in each time. For polyphonic instruments it may result that the instrument plays several pitches at each moment.

Due to the dynamic evolvement of the pitch over time, the filter design is dynamic,thus, the filter parameters must change according the content of the signal. This constraint leads us to adapt the Overlap Add (OLA) in the filtering process. We analyse the signal in overlapped frames where we modify the filter parameters in each frame according to the exact pitch value in the frame.

For sake of simplicity, we describe the filtering algorithm for a single window $\mathbf{x}$, where a single actual pitch within this window is known and equals $\hat{f}_0$. With the knowledge of $\hat{f}_0$ we model the samples in the window $\mathbf{x}$ by a harmonic model with additive WGN as in Equation (1). Using simple trigonometry identity we rewrite Equation (1) as:

$$x(n) = \sum_{k=1}^{K} a_k \cos(2\pi \hat{f}_0 k n) + b_k \sin(2\pi \hat{f}_0 k n) + w(n) \qquad (3)$$
$$n = 0, 1, \ldots, N-1$$

The unknown parameters in this model are the amplitudes of the sinusoids. We shall see that in order to filter the signal, we do not need to estimate explicitly the amplitude parameters. Let us denote

$$\mathbf{A_c} = \begin{pmatrix} 1 & \ldots & 1 \\ \cos(2\pi 2\hat{f}_0) & \ldots & \cos(2\pi K\hat{f}_0) \\ \vdots & \ldots & \vdots \\ \cos(2\pi \hat{f}_0(N-1)) & \ldots & \cos(2\pi \hat{f}_0 K(N-1)) \end{pmatrix} \qquad (4)$$

$$\mathbf{A_s} = \begin{pmatrix} 0 & \ldots, & 0 \\ \sin(2\pi 2\hat{f}_0) & \ldots & \sin(2\pi K\hat{f}_0) \\ \vdots & \ldots & \vdots \\ \sin(2\pi \hat{f}_0(N-1)) & \ldots & \sin(2\pi \hat{f}_0 K(N-1)) \end{pmatrix} \qquad (5)$$

The matrices A_c and A_s contain the sinusoidal components that form up the harmonic signal. We denote by $\mathbf{H}$ a matrix which is a concatenation of A_c and A_s.

$$\mathbf{H} = (A_c\ A_s) \qquad (6)$$

We now can represent the samples in each window $\mathbf{x}$ in terms of a linear model:

$$\mathbf{x} = \mathbf{H}\theta + \mathbf{w} \qquad (7)$$

where θ is the unknown amplitude vector $\theta = [a_1, \ldots, a_K, b_1, \ldots, b_K]^T$. It can be shown (Kay 1993) that the minimum variance unbiased estimator is

$$\hat{\theta} = (\mathbf{H^T H})^{-1} \mathbf{H^T x} \qquad (8)$$

Once we have the estimation of the unknown amplitudes of the sinusoids which form up the pitch, we use them to filter the desired pitch by:

$$\mathbf{y} = \mathbf{H}(\mathbf{H^T H})^{-1} \mathbf{H^T x} \qquad (9)$$

The rows of the matrix $\mathbf{H}(\mathbf{H^T H})^{-1}\mathbf{H^T}$ can be interpreted as FIR filter coefficients. The frequency response of the filter (See Fig. 1) shows that the filter passes $\hat{f}_0$ the fundamental frequency along with its partial in 0 dB, which means that these frequencies appear in the output signal in the same magnitude as in the input signal. All the other frequencies are suppressed by a factor that corresponds to the relative power between the main lobe and the sidelobes of the filter window. [1].

[1] In our application, the suppression of non-harmonic frequencies was

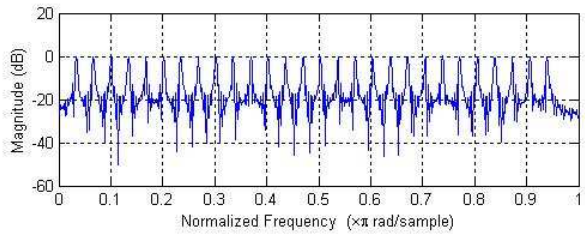

Figure 1: Frequency response of the harmonic filter

We emphasis that the model that we used is not complete. Harmonic model was used to model the pitch that we wish to filter and white gaussian noise was used to model the remainder of the spectrum. Using WGN might not be the correct model to the remainder of the spectrum, since it contains other instruments which might be harmonic or percussive and they exhibit high correlations between successive data samples[2]. However, since in this scenario the algorithm has information only about the one instrument that we wish to filter, the WGN assumption is the least committing one. Although incomplete, this model gives sufficiently good results.

2.3 Vocal Filtering

This algorithm can also be used in musical recordings with vocal singing such as Pop or Opera recordings. We distinguish between two different scenarios: the first is when we assume that we have only the score information of the singer. This can be for example a pop song with a rich accompaniment, where the accompaniment score information is not given. The second case is when we assume that we have both the score information of the singer and its accompaniment. This could be, for example, when there is light accompaniment such as a single piano, as in classical Lieder.

The filtering process of the vocal singing is similar to the filtering process of instruments. However, due to several unique characteristics of human voice we need to modify the algorithm. The filter algorithm which is based on the harmonic model can be applied only to the voiced parts of the singing. The unvoiced parts are noise like, and the harmonic model is inadequate. The alignment part already contains an algorithm for separating between voiced and unvoiced parts of the pitch. We process the voiced components using the algorithms that were described in sections (2.2) and (2.5). To process the unvoiced part we use a simple model in which we assume that the high frequencies in the spectrum of the unvoiced component belong to the singer and the low frequencies components belong to the accompaniment. Thus, we do high-pass filtering of the unvoiced part and the resulting signal is associated with the singer.

2.4 Subtracting an Instrument or Voice from a Recording

In many situations, especially when the score information is incomplete, it is useful to keep the accompaniment, whose score is unavailable to us, while suppressing the instrument or singer whose score we do have. For example, in karaoke, the recording contains the soloist along with the accompaniment, and the process removes the soloist part while keeping the accompaniment.

In order to solve this task we design the filter in the same way that was described in section (2.2), but we modify equation (9) to:

$$\mathbf{y} = \mathbf{x} - \mathbf{H}(\mathbf{H^T H})^{-1}\mathbf{H^T x} \tag{10}$$

In other words, we subtract the estimation of the soloist from the original signal. This gives an estimation to the accompaniment in the recording by projection of the recording on a subspace that is orthogonal or complementary to the signal space that described the soloist. All other details remain the same as in filtering the soloist scenario.

2.5 Adjusting the Balance between Several Instruments

The second scenario that we treat is balancing between several instruments in a musical piece. In this scenario, we have at our disposal score information about several instruments that are playing together. The input now contains the amount (measured in dB units) of boosting or suppression of each instrument. The filter design in this case differs from the filter design that was described above in two major points: First we have to extend the harmonic model from a single instrument to a group of instruments. Second, we must constrain the filter to have different magnitude responses which match the boosting/suprression request for the different instruments.

The extension of the harmonic model to group of different instruments is straight-forward. Assume that we have P instruments and let us denote in $\hat{f}_0^{(1)}, \hat{f}_0^{(2)} \dots \hat{f}_0^{(P)}$ the fundamental frequencies of all of these instruments . The extension to Equation (1) is then:

$$x(n) = \sum_{p=1}^{P}\sum_{k=1}^{K} a_k^{(p)} \cos\left(2\pi \hat{f}_0^{(p)} kn\right) + b_k^{(p)} \sin\left(2\pi \hat{f}_0^{(p)} kn\right) + w(n) \tag{11}$$

more than 10 dB. The reason for this number is the fact that the short-time analysis process can be viewed as multiplying the signal with a rectangular window. One of the characteristics of a rectangular window is that its frequency response resembles a sinc function, which has a difference between the main lobe to the first side-lobe of 13 dB. Better suppression of the non-harmonic components can be achieved using other types of windows.

[2]WGN assumes that the remaining part of the signal contains samples that are uncorrelated and Gauss distributed

In order to represent the signal in a linear model we use again the sinusoids matrix A_c and A_s. We denote now the matrix $\mathbf{H}$ as concatenation of all pairs of A_c and A_s

$$\mathbf{H} = (A_c^{(1)} A_s^{(1)} \dots A_c^{(P)} A_s^{(P)}) \tag{12}$$

The data can now be represented with a linear model as in (7) and the best linear estimator $\hat{\theta}$ to the amplitudes of all pitches is given by (8).

Once we have $\hat{\theta}$ we can filter the signal. The constraint that we have on the magnitude response of the filter is formulated as a gain matrix $\mathbf{G}$. $\mathbf{G}$ is diagonal matrix, which contains the magnitude response values for each pitch (fundamental and its partials). The filtering process is then defined by

$$\mathbf{y} = \mathbf{H}\mathbf{G}(\mathbf{H^T}\mathbf{H})^{-1}\mathbf{H^T}\mathbf{x} \tag{13}$$

3 Experimental Results

All examples are available at http://www.cs.huji.ac.il/∼chopin/VMix/index.html. The first example is an instrumental music recording - Mozart's violin sonata. It contains two instruments - violin and piano. Using the virtual mixer we tried to filter out the violin part. We assumed that we know the score information only for the violin part. We then used the algorithm that was described in section (2.2). As can be heard, the piano part in the modified signal is almost completely unhearable. The filter cut the piano part by 13 dB. For this example this is almost true isolation of the violin part. Figure 2 depicts this filtering process.

The second recording was a recording with a vocal part. We took a pop song - 'Summertime" sang by Ella Fitzgerald. As in the instrumental piece, we assumed that we know the score information of the singing. We then used the virtual mixer to modify the singer. As can be heard, the filtered voice contains both the voiced and unvoiced parts of the singing, which correspond to different processing algorithms used by the system. We then filtered out the accompaniment part using the algorithm described in section (2.4). With both the voice and the accompaniment, we virtually mixed the two parts with different balance between the soloist and its accompaniment.

The last example that we tested was a vocal recording with a strong percussion accompaniment. Since our filter cannot handle percussion instruments due to its assumption of the harmonic model, we wanted to test how good it handles the percussion instruments when they exists as accompaniment. The recording that we chose was the beetles song 'help" from which we chose a solo part with strong percussion accompaniment. We had the score information of the singer and using this information we filtered out the singer voice and the accompaniment. Then we re-mixed them again with different balance between the singer and the accompaniment. The result was quite good, the re-mixed sound preserve the strong percussion in the accompaniment.

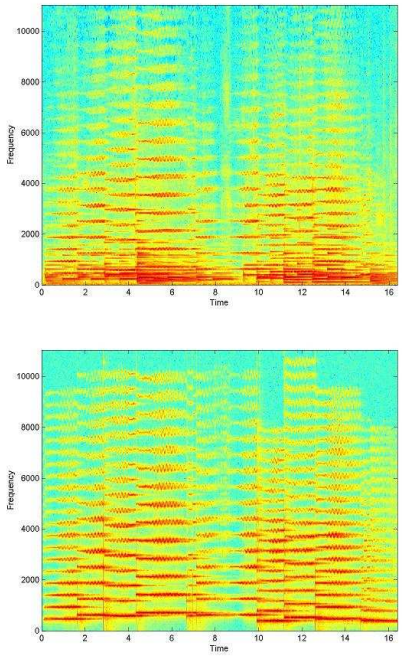

Figure 2: Filtering single instrument: The top figure is recording of piano and violin. Using the harmonic filter we extract the violin part (bottom figure)

References

Cook, M. P. and B. G. J (1994). *Separating simultaneous sound sources: issues, challenges, and models, Speech Recognition and Speech Synthesis*. John Wiley and Sons.

Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing*. Prantice Hall.

Roads, C., S. Pope, and A. Piccially (Eds.) (1997, June). *Musical Signal Processing*. Swets & Zeitlinger.

Rodet, R. (1997). Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models. In *IEEE Time-Frequency and Time-Scale Workshop*.

Serra, X. (1989, Oct.). *A System for sound analysis/transformation/synthesis based on deterministic plus stochastic decomposition*. Ph. D. thesis, Stanford University.

Shalev-Shwartz, S., S. Dubnov, N. Friedman, and Y. Singer (2002). Robust temporal and spectral modeling for query by melody. In *SIGIR*.

Therrien, C. W. (1992). *Discrete random signals and statistical signal processing*. Prentice Hall.

Virtanen, T. and A. Klapuri (2002). Separation of harmonic sounds using linear models for the overtone series. In *Proc. ICASSP 2002*.