# AN ANALYSIS PRIOR BASED DECOMPOSITION METHOD FOR AUDIO SIGNALS

*Omer Deniz Akyildiz, İlker Bayram**

Istanbul Technical University
Department of Electronics and Telecommunications Engineering
Maslak, 34469, Istanbul, Turkey

## ABSTRACT

We consider the problem of separating the tonal and the transient components of an audio signal. We employ mixed norms with different groups on the analysis coefficients to formulate the problem. We provide an algorithm for the solution of the problem and demonstrate that the formulation can be effective for the given task. We also provide a brief discussion on the difference with a synthesis prior based formulation.

***Index Terms***— Structured sparsity, analysis prior, mixed norms , morphological component analysis.

## 1. INTRODUCTION

We consider the problem of decomposition of an audio signal into tonal and transient parts. Tonal part consists of oscillatory components. In contrast, transient part includes impulsive components (like percussion). The proposed model assumes that considered signal is sum of tonal and transient parts.

Our approach is to model tonal and transient parts separately and to use these models to formulate the decomposition task as a minimization problem. We model the components in two stages. Firstly, we choose suitable Short-Time Fourier Transforms (STFT) to represent each component. Secondly, we use 'mixed norms with different groupings' on these STFT coefficients.

In the last decade, the problem of decomposition of signals into tonal and transient parts is studied in the context of regularized inverse problems. One of the well-known approaches to the signal decomposition problem is called Morphological Component Analysis (MCA) [1]. This decomposition method aims to separate two components of an image (also can be applied to audio) which can be sparsified by different dictionaries. In the MCA framework, the problem is formulated as a multilayer $\ell_1$ regularized inverse problem which is based on the analysis or synthesis approach. However, from a Bayesian perspective, the usage of $\ell_1$ norm as a signal coefficient prior puts into model an independence assumption between coefficients. $\ell_1$-regularization results in 'sparse' solutions but in an unstructured way. For audio signals, which are highly structured, $\ell_1$ regularized inverse problem formulations are not suitable. To obtain special structures such as tonal or transient part, which are sparse in frequency and time respectively, more structured estimation methods should be used. To overcome this problem, methods based on mixed norms as signal priors [2], [3], [4], [5] are studied in recent years. Multilayer decomposition methods based on mixed norm regularization are applied for audio processing [6]. In these works, formulations are based on the synthesis prior. In contrast, our formulation is based on analysis prior. (see [7] and [8] for differences and theoretical discussion).

*Outline*

In Section 2, we provide our problem formulation and discuss time-frequency distributions with their relations to mixed norms in detail. Then, in Section 3, we provide an algorithm to solve our problem. In Section 4, we briefly discuss the synthesis based decomposition formulation. In Section 5, we present our results and discuss the differences between analysis based approach and synthesis based approach. Section 6 is the conclusion.

## 2. PROBLEM FORMULATION

Given a mixture signal $\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are tonal and transient parts respectively, we formulate our problem as,

$$(\mathbf{x}_1^*, \mathbf{x}_2^*) = \underset{\mathbf{x}_1,\mathbf{x}_2}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y}-\mathbf{x}_1-\mathbf{x}_2\|_2^2+\lambda_1 \|\mathbf{x}_1\|_a+\lambda_2 \|\mathbf{x}_2\|_b. \quad (1)$$

where $\lambda_1$ and $\lambda_2$ are parameters of the model. In this formulation, we define norms, $\|.\|_a$ and $\|.\|_b$ such that,

(i) $\|.\|_a$ assumes low values for the tonal content,

(ii) $\|.\|_b$ assumes low values for the transient content.

We define $\|.\|_a$ and $\|.\|_b$ in two steps. First, we use two STFTs with different time-frequency resolutions. Second, we

*This work is supported by TUBITAK under project 110E240.

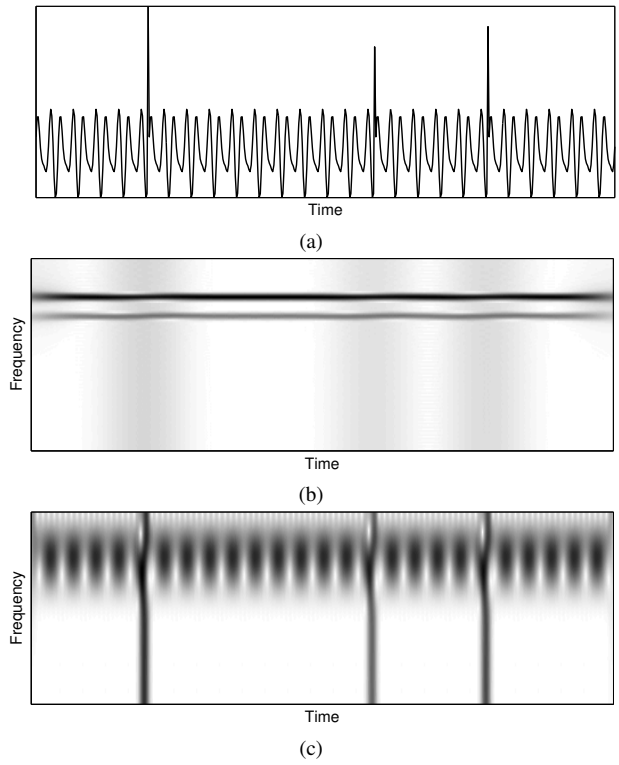


**Fig. 1**: (a) The mixture signal. This signal is sum of a oscillatory part that modeling the tonal part and clicks that modeling the transient part. In (b), STFT coefficients of the mixture signal with a long analysis window can be seen. In this panel, horizontal content is strong and sharper. In (c), STFT coefficients of the mixture signal with a short analysis window can be seen. In this panel, vertical content is strong and sharper.

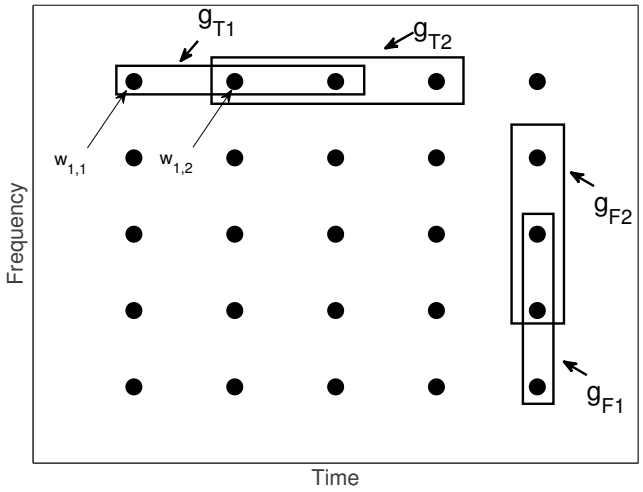


**Fig. 2**: Groups and overlaps in TF Plane. In this Figure, group size is 3 and groups are maximally overlapping.

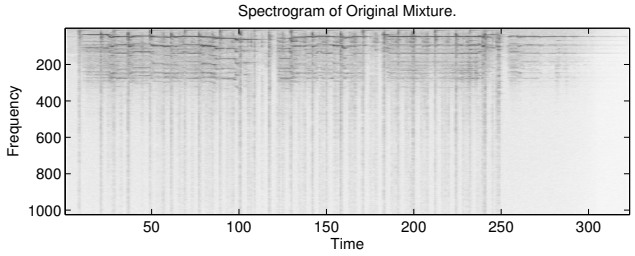


**Fig. 3**: Spectrogram of the mixture signal used in experiment with long analysis window.

define mixed norms to penalize particular structures in these two parametrized STFTs.

### 2.1. Parameters of the STFT

We employ STFTs with different time-frequency resolutions to analyze each component. The selection of the parameters is based on the observation that, in a time-frequency map, tonal components produce horizontal lines and transient components produce vertical lines. In order to further emphasize this difference, we employ an STFT with a high frequency resolution for the tonal component and an STFT with a high time-resolution for the transient component.

To gain intuition, consider the signal in Fig. 1(a). This signal consists of an oscillatory part and three clicks. We parametrize two STFTs to better represent each component. Fig. 1(b) shows the spectrogram of the signal when a long analysis window is used. We observe that, the horizontal lines are sharp wheras the vertical lines are rather diffuse. Similarly, an STFT with a short analysis window has a better time-resolution and is more suitable for representing the clicks. Fig. 1(c) shows the spectrogram of the signal when a short analysis window is used. In this representation, the vertical content is sharper than the horizontal content.

### 2.2. Mixed Norms with Different Groups

We use 'mixed norms with different groups' for each component. In order to clarify this, suppose that Fig. 2 depicts the STFT coefficients, where the coefficients are denoted by $w_{i,j}$ and the groups formed along time and frequency axes are denoted by $g_{T(i,k)}$ and $g_{F(i,k)}$ respectively. Here, we think of $g_{T(i,1)}$ as the collection $(w_{i,1}, w_{i,2}, w_{1,3})$ so that $\|g_{T(i,1)}\|_2 = \sqrt{|w_{i,1}|^2 + |w_{i,2}|^2 + |w_{i,3}|^2}$. In this setting, the norm defined by

$$\|w\| = \sum_{i,j} \|g_{T(i,j)}\|_2 \qquad (2)$$

assumes small values if the coefficients are clustered in horizontal groups – we refer to [3]. Similarly, if the norm is defined based on groups along the frequency axis (i.e. $g_F$) the norm will assume small values when the coefficients are clustered in vertical groups.

### 2.3. Mixed Norms of STFT Coefficients

In (1), we use the mixed norm of whole time-frequency representations with different groupings. This setting penalizes particular structures and leads to 'structured' solutions in the sense of time-frequency distributions. For the tonal part, we obtain a structured solution which has sharp horizontal structures without any vertical structure. Similarly, for the transient part, we obtain a structured solution which has sharp vertical structures without any horizontal structure.

In order to obtain the tonal part, we choose $\|.\|_a = \|A_1\mathbf{x}_1\|_{2,1}$ where $\|.\|_{2,1}$ is a mixed norm with groups formed along the time axis. $A_1$ is a transform with long analysis window. To obtain the transient part, we choose $\|.\|_b = \|A_2\mathbf{x}_2\|_{2,1}$ where $\|.\|_{2,1}$ is a mixed norm with groups formed along the frequency axis. $A_2$ is a transform with short analysis window. We discussed the properties of STFTs with long and short analysis windows in Section 2.1.

## 3. MINIMIZATION ALGORITHM

Recall that, our goal is to minimize,

$$J(\mathbf{x}_1,\mathbf{x}_2) = \frac{1}{2}\|\mathbf{y}-\mathbf{x}_1-\mathbf{x}_2\|_2^2 + \lambda_1\|\mathbf{x}_1\|_a + \lambda_2\|\mathbf{x}_2\|_b. \quad (3)$$

A coordinate-descent type algorithm [9] can be used for minimization. At each step, a typical coordinate-descent type algorithm fixes some variables. This means, by fixing some variables for each iteration, algorithm solves a particular optimization problem with respect to one variable. The method is summarized in Algorithm 1.

**Algorithm 1** Algorithm for minimizing (3).

1: **repeat**
2: $\quad\mathbf{x}_1 = \text{argmin}_u\ J(u,\mathbf{x}_2) \quad \{\text{P1}\}$
3: $\quad\mathbf{x}_2 = \text{argmin}_u\ J(\mathbf{x}_1,u) \quad \{\text{P2}\}$
4: **until** Convergence

In every iteration, P1 and P2 appear as new minimization problems. These problems can be regarded as mixed-norm denoising problems,

$$\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmin}}\ \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\|A\mathbf{x}\|_{2,1} \quad (4)$$

If we look our cost function, which is defined by (3), by assuming one component as constant, one can see that new minimization problem is equivalent to (4). When $A$ is the analysis operator of the frame, this problem has a different minimization procedure than iterative/shrinkage thresholding algorithms. This problem is studied in [10] and algorithms for minimization are provided.

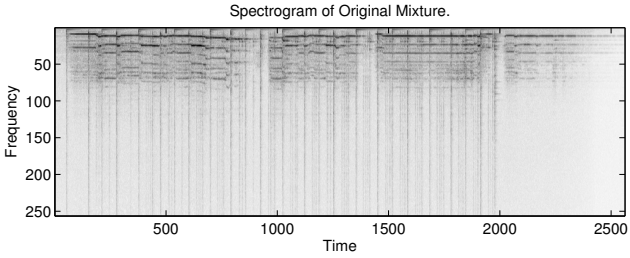


**Fig. 4**: Spectrogram of the mixture signal used in experiment with short analysis window.

## 4. DECOMPOSITION VIA SYNTHESIS PRIOR

In the next section, we will present decomposition results for synthesis prior based algorithm for comparison to analysis prior case. Let us now write the problem formulation for the synthesis approach.

With the synthesis prior, the decomposition problem can be formulated as,

$$\mathbf{w}_1^*, \mathbf{w}_2^* = \underset{\mathbf{w}_1,\mathbf{w}_2}{\text{argmin}}\ \|\mathbf{y}-S_1\mathbf{w}_1-S_2\mathbf{w}_2\|_2^2+\lambda_1\|\mathbf{w}_1\|_a+\lambda_2\|\mathbf{w}_2\|_b \quad (5)$$

This problem formulation is different from the analysis prior formulation and has been discussed in [2].

## 5. RESULTS AND COMPARISON

We present the decomposition of an audio signal which consists of tonal and transient parts. Audio signal is the sum of reed flute and darbuka (a kind of percussion). We try to model reed flute as tonal part and darbuka as transient part. We give the results obtained by both analysis and synthesis approaches for comparison.[1] We omit the mathematical details of the differences between analysis and synthesis approaches when frames are used (see [7] and [8]).

### 5.1. Experimental Setting

In previous sections, we proposed the decomposition algorithm via analysis prior based formulation. We compare this approach with synthesis approach. Before discussing results, we have to determine our parameters. We use same parameters for two approaches to see the similarities or differences.

We will discuss how to choose regularizer weights ($\lambda_i$'s), window sizes, group sizes, group shapes.

We choose regularizer weights ad hoc. From an intuitive point of view, regularizer weights adjust the suppression of particular components. However, it is not possible to easily

[1]The results for synthesis formulation are obtained by our implementation.

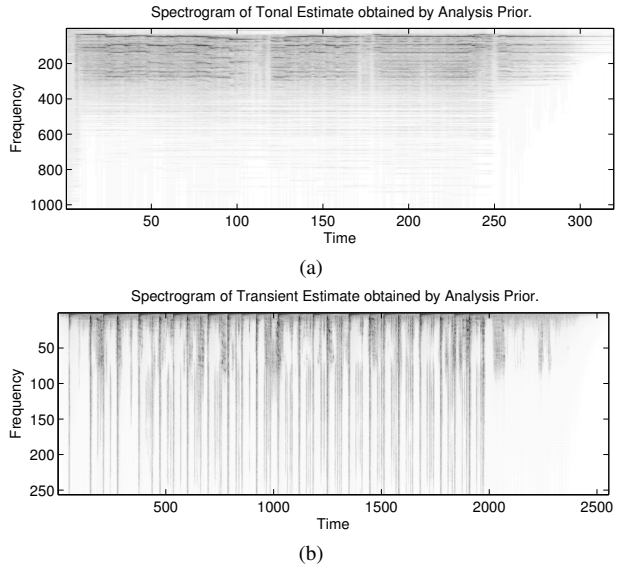


**Fig. 5**: (a) Spectrogram of the tonal estimate via analysis prior. The spectrogram mainly contains horizontal structure that represents tonal part. (b) Spectrogram of the transient estimate via analysis prior. The spectrogram mainly contains vertical structure that represents transient part.

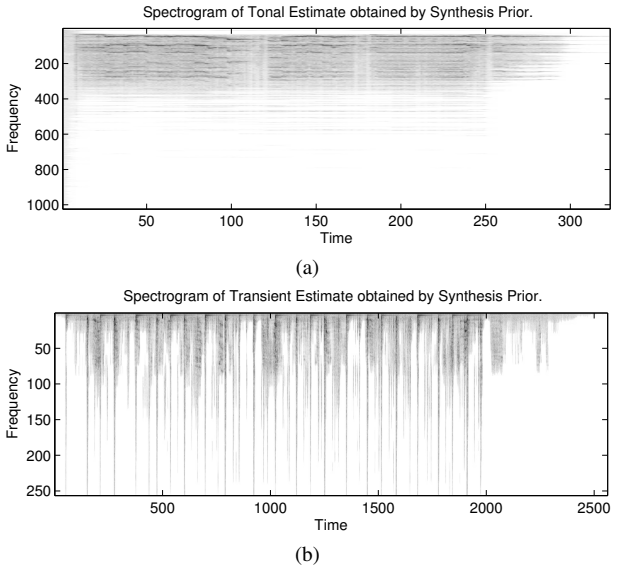


**Fig. 6**: (a) Spectrogram of the tonal estimate via synthesis prior. The spectrogram mainly contains horizontal structure that represents tonal part. (b) Spectrogram of the transient estimate via synthesis prior. The spectrogram mainly contains vertical structure that represents transient part.

adjust which component is suppressed because of the interaction between components through coordinate-descent algorithm. For regularizer weights, we choose $\lambda_1 = 0.6$ and $\lambda_2 = 0.3$. In several signals, regularizer weights should be selected properly to obtain perceptually meaningful estimates.

It is also important to choose proper STFT frames to represent audio signals. As we discussed before, we use long windows to represent tonal part and short windows to represent transient part. We used a smooth window of length $2048$ (the window is selected so as make the STFT a tight frame) and a Hop size of $1024$ samples. For the transient component, we use a similar smooth window of length $512$ and a Hop size of $128$ samples.

Group sizes and shapes are also important for us. In our experiment, we model the tonal part with groups formed by taking $15$ neighboring coefficients in the same subband with maximal overlap. For modeling transient part, we use a rectangular group shape with width $2$ and size $15$ with maximal overlap. We note that an interesting approach can be found in [6] which proposes to weight these groups with triangular windows.

The obtained spectrograms via analysis and synthesis approaches can be seen in Fig. 5 and Fig. 6 respectively. Also, SNR values of these estimates are tabulated in Table 1.

| Table 1 | | |
|---|---|---|
| SNR Values | Analysis | Synthesis |
| Tonal | 15.58 dB | 15.16 dB |
| Transient | 5.05 dB | 4.68 dB |

## 6. DISCUSSION AND CONCLUSION

We presented the results for analysis prior based decomposition method and compared it with the synthesis based approach.

Analysis prior based estimation results in 'smoother' spectrograms and have more non-zero coefficients than synthesis prior based estimations. This leads to a little difference in terms of perceptual quality. Analysis prior based estimates seem to be less contaminated by musical noise compared to the synthesis prior estimates.

Experiments presented in this paper can be listened from `http://web.itu.edu.tr/akyildizom/Thesis/decomposition1`.

## 7. ACKNOWLEDGEMENTS

O. D. Akyildiz thanks ITU MSPR Group for their support. Also, the authors thank Prof. B. Bozkurt, Bahcesehir University, Istanbul, for providing the data.

## 8. REFERENCES

[1] J. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Transactions on Image Processing*, vol. 14, pp. 1570–1582, (2005).

[2] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signal models with

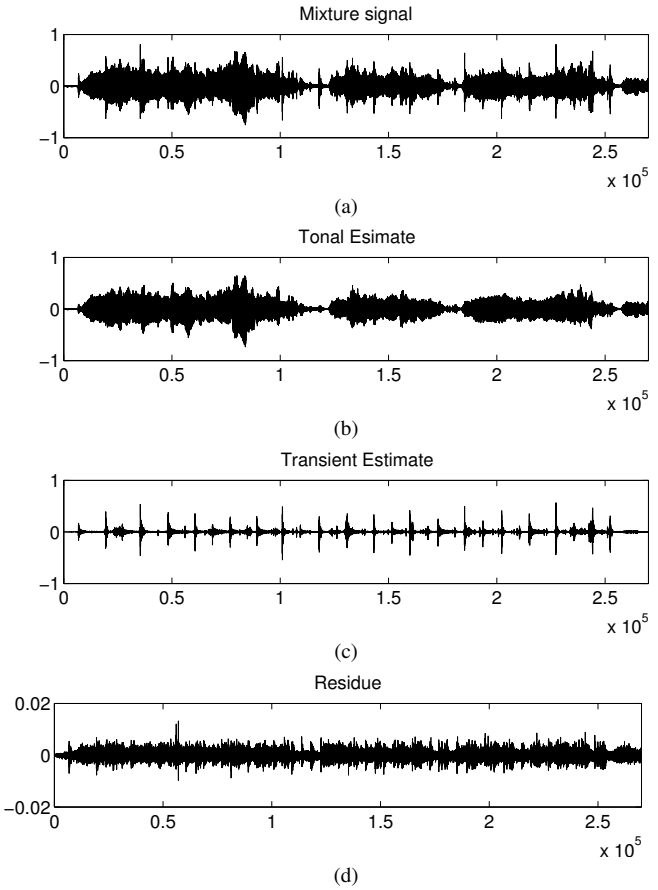


**Fig. 7**: (a) Mixture signal with energy 4129.9. (b) Tonal estimate obtained by analysis prior with energy 3554.07. (c) Transient estimate obtained by analysis prior with energy 243.77. (d) Residue with energy 0.31.

dependent coefficients," *Signal, Image and Video Processing*, vol. 3, no. 3, pp. 251–264, Sept. (2009).

[3] M. Kowalski, "Sparse regression using mixed norms," *Journal on Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303–324, Nov. (2009).

[4] Matthieu Kowalski and Bruno Torrésani, "Structured sparsity: From mixed norms to structured shrinkage," in *Signal Processing with Adaptive Sparse Structured Representations (SPARS '09)*, Saint Malo, France, (2009).

[5] İ. Bayram, "Mixed norms with overlapping groups as signal priors," in *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'11)*, (2011).

[6] K. Siedenburg and M. Dörfler, "Structured sparsity for audio signals," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx '11)*, (2011).

[7] Michael Elad, Peyman Milanfar, and Ron Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, pp. 947–968, (2007).

[8] I. W. Selesnick and M. A. T. Figueiredo, "Signal restoration with overcomplete wavelet transforms: Comparison of analysis and synthesis priors," *Proceedings of SPIE*, vol. 7446 (Wavelets XIII), August.

[9] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, (1984).

[10] İ. Bayram, "Denoising formulations based on support functions," İstanbul Technical University, Department of Electronics and Communications Engineering Technical Report, (2011).