SPARSE CHROMA ESTIMATION FOR HARMONIC AUDIO

Ted Kronvall, Maria Juhlin, Stefan Ingi Adalbjörnsson, and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University, Sweden.
email: {ted, juhlin, sia, aj}@maths.lth.se

ABSTRACT

This work treats the estimation of the chromagram for harmonic audio signals using a block sparse reconstruction framework. Chroma has been used for decades as a key tool in audio analysis, and is typically formed using a Fourier-based framework that maps the fundamental frequency of a musical tone to its corresponding chroma. Such an approach often leads to problems with tone ambiguity, which we avoid by taking into account the harmonic structure and perceptional attributes in music. The performance of the proposed method is evaluated using real audio files, clearly showing preferable performance as compared to other commonly used methods.

Index Terms— chromagram, block sparsity, total variation, convex optimization, ADMM

1. INTRODUCTION

Signal processing for audio and music has during the recent decades experienced a great surge of development, especially in symbiosis with portable devices and smart phones, offering solutions for most kinds of audio applications, ranging from automatic chord transcription and cover song detection to advanced recommendation systems based on musical similarity (see, e.g., [1–4]). The vast scope of music genres, ranging over Western pop music and classical big band orchestras to Arabic folk music, makes general music signal processing a daunting task, and as a consequence the scope of the applications is usually restricted to a certain type of music, most often Western pop music. When categorizing music, there are a number of features one may choose to study, where one such feature is the chroma, which places each tone on a cyclic scale in order to mirror the perception of the human ear. The appeal of the chroma for classification is based on that two tones with substantially different frequency content will sound similar to the human ear [5]. This implies that any method that classifies audio signals without considering such psychoacoustical attributes will generally not classify two such tones as being similar. Today, many algorithms use chroma estimation as a sub-step to yield a crude estimate of the signal information, often constituting a key component in applications such as, e.g., cover song identification and audio thumbnailing (see, e.g., [6–8]). Typically, one is often interested in estimating the chromagram, which is a representation of a signal's chroma content over time. There are different ways of constructing the chromagram, but the basic steps often involve a pitch estimation, followed by a mapping to a chroma, which may be impeded by the largely overlapping spectral content of the chromas. Examples of such estimators are the one by Ellis [9], which uses a windowed short-time Fourier transform, and the one by Müller and Ewert [10], which uses a filterbank approach. In this work, we propose a further alternative, forming the chroma estimate directly from the audio signal using sparse modeling, adapting the sparse and block sparse signal representation frameworks introduced in [11] and [12], to account for the chroma structure. Thus, the signal is represented using a dictionary of candidate chromas, wherein all candidate octaves of the musical tone are represented as sets of harmonically related sinusoids, i.e., as a pitch signal, although without assuming any prior knowledge of the number of tones present in the signal, nor of the number of harmonics. To promote a chroma sparse solution without misclassifications, we introduce a specific chroma penalty term, which promotes tones with the expected harmonic spectral content over tones having only a spurious subset of the estimated harmonics. As will be shown, this may be accomplished by minimizing a sum of convex penalty functions that together will promote the sought chroma structure, mitigating the inherent ambiguities in the chroma representation.

This work was supported in part by the Swedish Research Council, Carl Trygger's foundation, and the Royal Physiographic Society in Lund.

Algorithm 1 The proposed CEBS algorithm

Initiate $\mathbf{z} = \mathbf{z}(0), \mathbf{u} = \mathbf{u}(0)$, and $\ell := 0$
repeat
  $\mathbf{z}(\ell) = \left(\mathbf{G}^H\mathbf{G}\right)^{-1}\mathbf{G}^H\left(\mathbf{u}(\ell) + \mathbf{d}(\ell)\right)$
  $\mathbf{u}^{(1)}(\ell+1) = \frac{\mathbf{y} - \mu\left(\mathbf{W}\mathbf{z}(\ell+1) - \mathbf{d}^{(1)}(\ell)\right)}{1+\mu}$
  $\mathbf{u}^{(2)}(\ell+1) = \bar{\Phi}\left(\Psi\left(\mathbf{z}(\ell+1) - \mathbf{d}^{(2)}(\ell), \frac{\lambda_2}{\mu}\right), \frac{\lambda_3\sqrt{M}}{\mu\sqrt{12}}\right)$
  $\mathbf{u}^{(3)}(\ell+1) = \Phi\left(\mathbf{F}\mathbf{z}(\ell+1) - \mathbf{d}^{(3)}(\ell), \frac{\lambda_3\sqrt{M}}{\mu\sqrt{12}}\right)$
  $\mathbf{d}(\ell+1) = \mathbf{d}(\ell) - \left(\mathbf{G}\mathbf{z}(\ell+1) - \mathbf{u}(\ell+1)\right)$
  $\ell \leftarrow \ell + 1$
until convergence

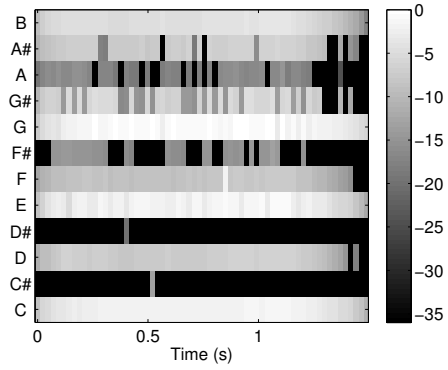

Fig. 1. The Ellis log-chromagram for the chord signal.

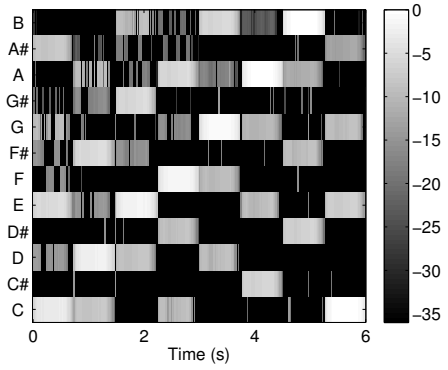

Fig. 2. The Ellis log-chromagram for the scale signal.

2. THE HARMONIC SIGNAL MODEL

Consider a noise-free tonal audio signal, such that the complex-valued[1] signal may be modeled as a sum of K distinct pitch signals, each consisting of L_k harmonically related sinusoids, i.e.,

$$x(t) = \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} a_{k,\ell} e^{i2\pi f_k \ell t} \tag{1}$$

for $t = 1, ..., N$, where $a_{k,\ell}$ denotes the amplitude of the ℓ:th harmonic in the k:th pitch, and with f_k and L_k denoting the normalized fundamental frequency and the number of sinusoids of the k:th source, respectively. The choice of model is motivated by the fact that most musical instruments produce sounds which may be well described as a harmonic series of sinusoids, having integer multiples of a fundamental frequency. Also, the data is modeled in the time domain, as this is shown to render more efficient estimates than using the magnitude short-time fourier transform (STFT) [14]. Because of the harmonic structure, the human perceptory system does not perceive the frequencies as being separate, but rather as a single musical tone. Furthermore, two tones having fundamental frequencies at a ratio of 2:1 are perceived as quite similar, referred to as being octave equivalent. A partial reason for this is that the harmonics of a tone with fundamental frequency $2f_0$, i.e., $2f_0, 4f_0, 6f_0, \ldots$, form a perfect subset of the harmonics for a tone one octave below. Most Western music use a cyclic scale divided into twelve semitones within each octave, spaced by a relative absolute frequency of $2^{1/12}$. These twelve tones are not only perceived as distinctly different from each other by our auditory system, but also as equally spaced, giving credit to the idea that our hearing is $\log$ tempered. The set of all tones being octave equivalent is called a chroma, with the twelve chromas being

$$C, C\#, D, D\#, E, F, F\#, G, G\#, A, A\#, \text{ and } B \tag{2}$$

[1] In order to simplify notation, we here examine the discrete-time analytic signal version (see, e.g., [13, 14]) of the measured audio signal.

and where the concatenation of a chroma with its octave number, e.g., A4, forms a musical tone. The fundamental frequency of a tone is thus detailed as

$$f_k = f_{\text{base}} \cdot 2^{c_k/12 + o_k} \tag{3}$$

where c_k and o_k are the chroma and octave of pitch k, respectively, and f_{base} denotes a normalized tuning parameter [2]. As a result, not only do the octaves within a chroma have coinciding harmonics, but so do tones with fundamental frequencies at other ratios, e.g., 3:2 (or nearly so) which is known as a "perfect fifth". Fifths are spaced by seven semitones and are commonly used together in musical compositions, as the overlapping spectral content is perceptually pleasant to hear. To account for this spectral ambiguity of the musicologic system, we propose to approximate (1) by an underdetermined signal model based on both chromas and octaves, taking into account the harmonic structure of the musical tone, such that

$$x(t) \approx \sum_{c=0}^{11} \sum_{o=O_{\min}}^{O_{\max}} \sum_{\ell=1}^{L_{\max}} a_{c,o,\ell} e^{i2\pi f_{\text{base}} 2^{(c/12+o)} \ell t} \tag{4}$$

where $c = 0, \ldots, 11$ are the twelve semitones ordered as in (2), and where $[O_{\min}, \ldots, O_{\max}]$ indicates the range of octaves considered. Furthermore, $L_{\max}$ is the maximal number of harmonics considered, and $a_{c,o,\ell}$ is the (complex-valued) amplitude for harmonic ℓ in the musical tone c, o. From (4), it is clear that the spectral content is discretized into $M = 12(O_{\max} - O_{\min})L_{\max}$ feasible frequencies, grouped within each octave and chroma. And, as noted above, many of the harmonics between chromas typically coincide, we deem it insufficient to simply map individual frequencies to chromas, as they will likely then also map to several other chromas, too. Instead, we propose the simultaneous mapping of the whole set of the tone's harmonics to a chroma. To this end, let

$$\mathbf{\Psi} = \left\{ \left\{ a_{c,o,1}, \ldots, a_{c,o,L_{\max}} \right\}_{o=O_{\min},\ldots,O_{\max}} \right\}_{c=0,\ldots,11} \tag{5}$$

be the structure of amplitudes for all possible frequencies in the chroma model in (4). As the set $\mathbf{\Psi}$ is much larger than

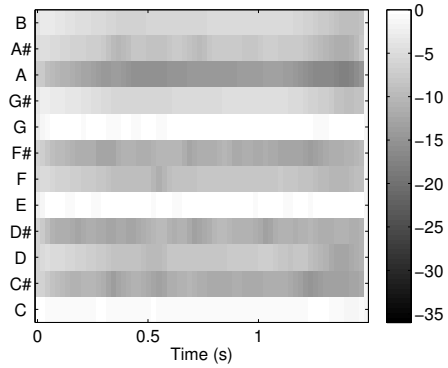

Fig. 3. The Müller and Ewert log-chromagram for the chord.

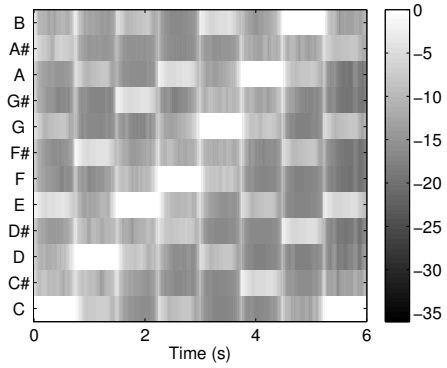

Fig. 4. The Müller and Ewert log-chromagram for the scale.

the actual solution set, most amplitudes, $a_{c,o,\ell}$, in (5) will be equal to zero, i.e., $\mathbf{\Psi}$ is sparse. If, for instance, only the key C#5 is played, then all amplitudes, except $a_{1,5,\ell}$, for those ℓ present in this tone, will be zero.

3. SPARSE CHROMA ESTIMATION

We proceed to detail the proposed chroma estimation procedure, which is formed without any prior knowledge of the number of tones present in the signal, nor the number of harmonics in each tone. Denoting the measured (noise-corrupted) signal $y(t)$, we strive to form the minimization $\min_{\mathbf{\Psi}} g_1(\mathbf{\Psi})$, where the squared model residual $g_1(\mathbf{\Psi})$ is defined as

$$g_1(\mathbf{\Psi}) = \sum_{t=1}^{N} \left| y(t) - \sum_{c=0}^{11} \sum_{o=O_{\min}}^{O_{\max}} \sum_{\ell=1}^{L_{\max}} a_{c,o,\ell} e^{i2\pi f_{\text{base}} 2^{(c/12+o)} \ell t} \right|^2 \tag{6}$$

As the number of harmonics in each tone, L_k, is unknown, we here select $L_{\max}$ such that $L_{\max} > \max_k\{L_k\}$, i.e., large enough. Clearly, such a minimization will not enforce the assumed sparsity of the signal, and we therefore impose constraints to ensure the sparsity of the solution. As noted in [11], such a sparse solution may be enforced using the ℓ_1-norm, such that amplitudes corresponding to low spectral power are set to zero. We thus extend the minimization with a constraint on

$$g_2(\mathbf{\Psi}) = \sum_{c=0}^{11} \sum_{o=O_{\min}}^{O_{\max}} \sum_{\ell=1}^{L_{\max}} |a_{c,o,\ell}| \tag{7}$$

which will penalize solutions with many components, thereby restricting the overall number of estimated amplitudes. Furthermore, due to the presence of the tones's harmonic structure, the signal will exhibit a strong group sparsity. To account for this behavior, reminiscent to [12, 15], we further introduce the penalty

$$g_3(\mathbf{\Psi}) = \sum_{c=0}^{11} \sqrt{\sum_{o=O_{\min}}^{O_{\max}} \sum_{\ell=1}^{L_{\max}} a_{c,o,\ell}^2} \tag{8}$$

which will enforce sparsity for the entire set of amplitudes within each chroma. It will thus allow us to map the spectral content of the signal to the most appropriate chroma, nulling the contribution of ambiguous chromas, despite their partially overlapping spectral content.

$$g_4(\mathbf{\Psi}) = \sum_{c=0}^{11} \sqrt{\sum_{o=O_{\min}}^{O_{\max}} \sum_{\ell=1}^{L_{\max}-1} |a_{c,o,\ell+1} - a_{c,o,\ell}|} \tag{9}$$

As formed in (9), the total variation penalty will also penalize non-zero amplitudes at wrong octaves within the chromas, ensuring an even sparser solution. Thus, in summary, we propose to estimate the chomas present in the observed signal by the (convex) minimization

$$\hat{\mathbf{\Psi}} = \arg\min_{\mathbf{\Psi}} \sum_{i=1}^{4} \lambda_i g_i(\mathbf{\Psi}) \tag{10}$$

where $\lambda_1 = 1$, and λ_i, for $i = 2, 3, 4$, are user-defined sparse regularizers which weigh the importance between each penalty function and the squared residual. To simplify notation, let

$$\mathbf{y} = \begin{bmatrix} y(1) & \dots & y(N) \end{bmatrix}^T \tag{11}$$

$$= \sum_{c=0}^{11} \mathbf{W}_c \, \mathbf{a}_c + \mathbf{e} \triangleq \mathbf{W}\mathbf{a} + \mathbf{e} \tag{12}$$

where $(\cdot)^T$ denotes the transpose, and

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_0 & \dots & \mathbf{W}_{11} \end{bmatrix}^T \tag{13}$$

$$\mathbf{W}_c = \begin{bmatrix} \mathbf{w}_c^{O_{\min}} & \dots & \mathbf{w}_c^{O_{\max}} \end{bmatrix}^T \tag{14}$$

$$\mathbf{w}_c = \begin{bmatrix} \mathbf{z}^1 & \dots & \mathbf{z}^{L_{\max}} \end{bmatrix}^T \tag{15}$$

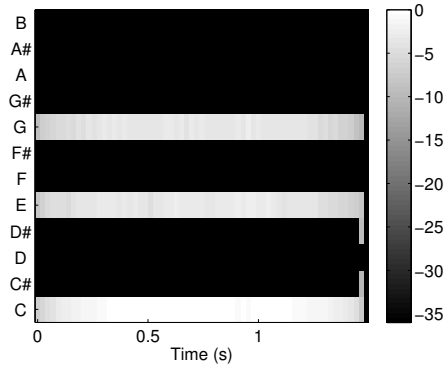

Fig. 5. The CEBS log-chromagram for the chord signal.

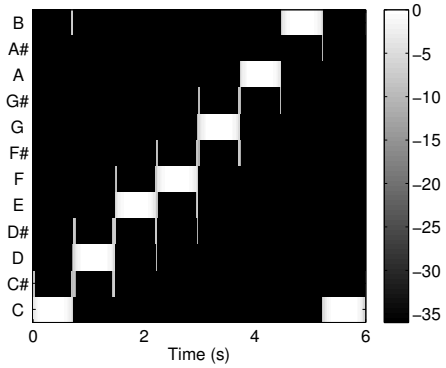

Fig. 6. The CEBS log-chromagram for the scale signal.

$$\mathbf{z}_c = \left[\begin{array}{ccc} e^{i2\pi 2^{c/12}1} & \dots & e^{i2\pi 2^{c/12}N} \end{array} \right]^T \quad (16)$$

$$\mathbf{a}_c = \left[\begin{array}{ccc} \mathbf{a}_{c,O_{\min}}^T & \dots & \mathbf{a}_{c,O_{\max}}^T \end{array} \right]^T \quad (17)$$

$$\mathbf{a}_{c,o} = \left[\begin{array}{ccc} a_{c,o,1} & \dots & a_{c,o,L_{\max}} \end{array} \right]^T \quad (18)$$

Thus, the block dictionary $\mathbf{W} \in \mathbb{C}^{N \times M}$, where M denotes the number of possible frequencies, has twelve blocks of chroma, such that each chroma is a block of $(O_{\max} - O_{\min})$ tones, with every tone, in turn, a block of $L_{\max}$ column Fourier vectors. This allows (10) to be expressed as

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \sum_{i=1}^{4} g_i(\mathbf{H}_i \mathbf{a}, \lambda_i) \quad (19)$$

where $\mathbf{H}_1 = \mathbf{W}$, $\mathbf{H}_2 = \mathbf{H}_3 = \mathbf{I}$, $\mathbf{H}_4 = \mathbf{F}$, and

$$g_1(\mathbf{W}\mathbf{a}, 1) = ||\mathbf{y} - \mathbf{W}\mathbf{a}||_2^2 \quad (20)$$

$$g_2(\mathbf{a}, \lambda_2) = \lambda_2 ||\mathbf{a}||_1 \quad (21)$$

$$g_3(\mathbf{a}, \lambda_3) = \lambda_3 \sum_{c=0}^{11} ||\mathbf{a}_c||_2 \quad (22)$$

$$g_4(\mathbf{F}\mathbf{a}, \lambda_4) = \lambda_4 ||\mathbf{F}\mathbf{a}||_1 \quad (23)$$

where $\mathbf{I}$ denotes the identity matrix, and where $\mathbf{F}$ is the first order difference matrix, having elements $\mathbf{F}_{i,i} = 1$ and $\mathbf{F}_{i,i+1} = -1$, for $1, \dots, M/12 - 1$, and zeros elsewhere. As the minimization in (19) is convex, it may be solved using one of the freely available interior-point based methods, such as SeDuMi [16] or SDPT3 [17], although it may be noted that these will scale poorly with increasing data length. Inspired by the results in [12], we here introduce a computationally efficient implementation based on the alternating direction of multipliers method (ADMM), see e.g., [17]. In broad terms, the ADMM framework can be used to solve convex optimization problems composed of a sum of two convex functions by introducing an auxiliary variable $\mathbf{u}$ so that the optimization can be split into two simpler independent sub-problems that are solved in an iterative fashion. Using the extension to a sum of several convex functions (see also [18]), we propose the Chroma Estimation using Block Sparsity (CEBS) algorithm, as given in Algorithm 1, where μ is an inner convergence variable, $\mathbf{d}$ a dual variable, and

$$\mathbf{G} = \left[\begin{array}{ccc} \mathbf{W}, & \mathbf{I}, & \mathbf{F} \end{array} \right] \quad (24)$$

$$\mathbf{u} = \left[\begin{array}{ccc} \mathbf{u}^{(2)T}, & \mathbf{u}^{(3)T}, & \mathbf{u}^{(3)T} \end{array} \right]^T \quad (25)$$

$$\mathbf{d} = \left[\begin{array}{ccc} \mathbf{d}^{(2)T}, & \mathbf{d}^{(3)T}, & \mathbf{d}^{(3)T} \end{array} \right]^T \quad (26)$$

$$\Phi(\mathbf{x}, \xi) = \frac{\max(|\mathbf{x}| - \xi, 0)}{\max(|\mathbf{x}| - \xi, 0) + \xi} \odot \mathbf{x} \quad (27)$$

$$\bar{\Phi}(\mathbf{x}, \xi) = \frac{\max(||\mathbf{x}||_2 - \xi, 0)}{\max(||\mathbf{x}||_2 - \xi, 0) + \xi} \mathbf{x} \quad (28)$$

such that the solution is given as $\hat{\mathbf{a}} = \mathbf{z}(\ell_{\text{end}})$.

4. NUMERICAL RESULTS

We proceed to examine the performance of the proposed algorithm, comparing with the (publicly available) estimators in [9, 10], using two audio signals from [19], namely a two channels FM-violin playing a middle C scale (all tones from C4 to C5) and a C-major chord, both in equal temperament, sampled at $f_s = 22050$ Hz, mixed to a single channel using the method in [10]. Figures 1-6 illustrate the resulting log-chromagrams for the Ellis, the Müller and Ewert, and the CEBS estimators. We have here divided the signal in segments of length $N = 1024$ samples (about 46 ms), having an overlap of 50%. For CEBS, we set $\lambda_2 = 0.05$, $\lambda_3 = 2.3$, and $\lambda_4 = 0.1$, which are chosen using some simple heuristics from the FFT (see, e.g., [12]). The tuning frequency is here set to $f_{\text{base}} = 440$, and results remains quite unchanged at ± 3 Hz. A more thorough sensitivity analysis of this parameter is beyond the scope of this work. Clearly, the CEBS estimator yields a preferable estimate, suffering from noticeably less leakage and spurious estimates. As can be seen in Figure 6, the algorithm may make some misclassifications at the onset of a new tone, due to the non-stationary nature and weakness of the signal in these frames; this may easily be mitigated by adding a post-processing with a lowpass filter.

5. REFERENCES

[1] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal Processing for Music Analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1088–1110, 2011.

[3] G. Dror, N. Koenigstein, and Y. Koren, "Web-Scale Media Recommendation Systems," *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2722–2736, Sept. 2012.

[4] T. Hofmann, "Latent Semantic Models for Collaborative Filtering," *ACM Trans. Inf. Sys.*, vol. 22, no. 1, pp. 89–115, Jan. 2004.

[5] R. Shepard, "Circularity in Judgements of Relative Pitch," *Journal of Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, Dec. 1964.

[6] M. A. Bartsch and G. H. Wakefield, "Audio Thumbnailing of Popular Music Using Chroma-based Representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, Feb. 2005.

[7] S. Kim and S. Narayanan, "Dynamic Chroma Feature Vectors with Applications to Cover Song Identification.," in *10th IEEE Workshop on Multimedia Signal Processing*, 2008, pp. 984–987.

[8] T.-M. Chang, E.-T. Chen, C.-B. Hsieh, and P.-C. Chang, "Cover Song Identification with Direct Chroma Feature Extraction from AAC Files," in *IEEE 2nd Global Conference on Consumer Electronics*, Oct. 2013, pp. 55–56.

[9] D. P. W. Ellis, "Chroma Feature Analysis and Synthesis," `http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/`, accessed Sept. 2014.

[10] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB Implementations for Extracting Variants of Chroma-based Audio Features," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011.*, 2011.

[11] J. J. Fuchs, "On the Use of Sparse Representations in the Identification of Line Spectra," in *17th World Congress IFAC*, Seoul, jul 2008, pp. 10225–10229.

[12] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-Pitch Estimation Exploiting Block Sparsity," *Elsevier Signal Processing*, vol. 109, pp. 236–247, April 2015.

[13] S. L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, September 1999.

[14] M. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Morgan & Claypool, 2009.

[15] X. Lv, G. Bi, and C. Wan, "The Group Lasso for Stable Recovery of Block-Sparse Signal Representations," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, 2011.

[16] R. H. Tutuncu, K. C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Ser. B*, vol. 95, pp. 189–217, 2003.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[18] M. A. T. Figueiredo and J. M. Bioucas-Dias, "Algorithms for imaging inverse problems under sparsity regularization," in *Proc. 3rd Int. Workshop on Cognitive Information Processing*, May 2012, pp. 1–6.

[19] M. Romain, "Sound examples," `https://ccrma.stanford.edu/~mromaine/220a/fp/sound-examples.html`, accessed Sept. 2014.