AN ITERATIVE FILTERBANK APPROACH FOR EXTRACTING SINUSOIDAL PARAMETERS FROM QUASI-HARMONIC SOUNDS

Harvey D. Thornburg, Randal J. Leistikow

Stanford University, CCRMA
660 Lomita Drive
Stanford, CA 94305

ABSTRACT

We propose an iterative filterbank method for tracking the parameters of exponentially damped sinusoidal components of quasi-harmonic sounds. The quasi-harmonic criteria specialize our analysis to a wide variety of acoustic instrument recordings while allowing for inharmonicity. The filterbank splits the recorded signal into subbands, one per harmonic, in which time-varying parameters of multiple closely-spaced sinusoids are estimated using a Steiglitz-McBride/Kalman approach. Averaged instantaneous frequency estimates are used to update the center frequencies and bandwidths of the subband filters; by so doing, the filterbank progressively adapts to the inharmonicity structure of a source recording.

1. INTRODUCTION

Our goal is to estimate all parameters of the following exponentially-damped sinusoidal model as accurately as possible, given the signal y_t.

$$y_t \quad = \quad \sum_{k=1}^{p} A_{k,t} e^{-\gamma_k t} \cos(\omega_k t + \phi_{k,t}), \quad t = 1 \dots N \quad (1)$$

Many cases arise where the signal is *quasi-harmonic*, i.e. there exists a *known* fundamental pitch ω_0 and integer $l_k > 0$ for which each $\omega_k \approx l_k \omega_0$ over the signal's duration. In (1), the frequencies, and also the decay rates, $\{\gamma_k\}$, are assumed constant; however, the phases $\{\phi_{k,t}\}$ and amplitudes $\{A_{k,t}\}$ may vary with time. The latter variations are useful in modeling local transient effects, since fast phase and amplitude variations may proxy for small, local deviations in frequencies and decay rates.

Classical transform-based methods such as the phase vocoder [5] and SMS [7], implicitly via the STFT, dissect the signal into uniformly-spaced subbands. Within each subband, the methods estimate time-varying parameters of one or more sinusoids. However, these raw methods fail to account for apriori information concerning fundamental pitch, harmonic structure, or additional information from an acoustical model. For instance, the subband centers, fixed by transform length, are often unrelated to the fundamental. Furthermore, some acoustic instruments, such as piano, exhibit a hierarchical spectral structure involving a fixed number of exponentially decaying sinusoids closely surrounding each harmonic as induced by coupling, e.g. between vertical modes of multiple strings and/or between vertical and horizontal modes on a single string [12]. As the overall system is linear, there exist *equivalent modes* which do superpose [4], yielding the familiar beating effects, multistage decay behavior [12], etc. Equivalent modes may be arbitrarily close in frequency. As such, STFT-based analysis fails to guarantee each subband preserves the grouping of modes for a given partial, or that individual modes can be fully resolved.

Classical pitch-synchronous methods (cf. [6]) fit a harmonic signal model, then extract sinusoidal parameters. However, the assumption of an exact harmonic structure may not be valid for all quasi-harmonic sounds.

On the other hand, acoustical model-based methods do explicitly incorporate apriori information about the source. For instance, the waveguide model of Aramaki et al. [1] fully reproduces the interaction of coupled strings in piano tones. Nonetheless, a drawback of purely acoustic-based methods is that the palette of resynthesis transformations restricts to that of the acoustical model's parameter space. Sinusoidal models, by contrast, allow the user to explore explicitly "non-physical" modifications while preserving the timbral richness of the acoustic source.

In an attempt to overcome the drawbacks of previously existing methods, we propose an analysis framework which makes flexible use of signal-specific knowledge, especially knowledge pertaining to harmonic structure, during the analysis, to the extent this knowledge is present. To this end, our framework employs an iterative filterbank which adapts to the signal's harmonic structure. The filterbank isolates each group of sinusoids surrounding a given partial, and uses the subsequent sinusoidal parameter estimates to refine the positioning and bandwidths of the component filters.

2. ITERATIVE FILTERBANK

2.1. Overview

A bank of bandpass filters splits the signal into subbands, one subband per harmonic. Multiple sinusoids may be associated with each harmonic. To manifest this hierarchy, we restructure (1) as follows:

$$y_t = \sum_{l=1}^{p} \sum_{k=1}^{p_l} A_{k,l,t} e^{-\gamma_{k,l} t} \cos(\omega_{k,l} t + \phi_{k,l,t}), \quad t = 1 \dots N \quad (2)$$

Here l is the harmonic number, p is the number of harmonics, and p_l is the number of sinusoids surrounding the l^{th} harmonic.

For each iteration, we estimate $\omega_{k,l}$ and track $\phi_{k,l,t}$, then extract for each l, an average instantaneous frequency estimate for all k and t:

$$\hat{\omega}_l(i) \quad = \quad \frac{1}{p_l} \sum_k \left[\omega_{k,l} + \frac{1}{N-1} \left(\bar{\phi}_{k,l,N} - \bar{\phi}_{k,l,1} \right) \right] \quad (3)$$

Here i is the iteration number and $\bar{\phi}_{k,l,t}$ is the phase *deviation* subsequent to *unwrapping* the instantaneous phase $\omega_{k,l} t + \phi_{k,l,t}$.

At the conclusion of each iteration, the l^{th} center frequency is replaced by the instantaneous frequency estimate.

Absent instrument-specific information, the center frequencies are initialized as harmonic, i.e. $\{\omega_{c,l}(0)\}_{l=1}^{p} = l\omega_0$,where $\omega_{c,l}(i)$ is the l^{th} center frequency. If an acoustic model is known, however, an ideal inharmonicity profile may be used instead. An example inharmonicity profile from Bensa's piano model [2] follows.

$$\omega_{c,l}(0) \quad = \quad \sqrt{-\left(b_1 + b_2' l\omega_0^2\right)^2 + \omega_0^2(l^2 + Bl^4)} \qquad (4)$$

Here b_1, and $b_2^{'}$, are respectively the global loss and normalized frequency-dependent loss coefficient, and B is the normalized stiffness characteristic [3] of a piano string.

Each filter bandwidth should suffice to encompass the frequency spread of all sinusoids surrounding a given partial, as well as the center frequency deviation due to uncertainty about the actual harmonic structure. The latter uncertainty decreases as iterations progress. Hence, each filter bandwidth may progressively narrow to encompass only the expected frequency spread.

2.2. Optimal filter design

Each bandpass filter is designed as zero phase by a modified linear programming approach inspired by [10]. Various optimality conditions on the frequency response $H(\omega)$ are satisfied at N_g select frequency grid points $\{\omega_{g,j}\}_{i=1}^{N_g} \in [0, \pi]$, Fig. 1 illustrates a sample desired response for the filter surrounding the second harmonic, and details of its subband response. Vertical dashed lines are aligned with the center frequencies of the filters.

Two types of optimality conditions exist. *Hard* constraints must be met exactly; e.g.:

1. *Passband concavity*: $\partial^2 H(\omega_{g,j})/\partial\omega_{g,j}^2 \geq 0$ eliminates ripples in the passband.
2. *Transition width monotonicity*: $\partial H(\omega_{g,j})/\partial\omega_{g,j} \geq 0$ for $\omega_{g,j}$ in the lower transition region; $\partial H(\omega_{g,j})/\partial\omega_{g,j} \leq 0$ for $\omega_{g,j}$ in the upper region.

Soft constraints, on the other hand, are met within a tolerance optimized globally for all constraints. The goal is to minimize t, such that each constraint is satisfied within $c_j t$, where c_j^{-1} is the *importance weight* for the j^{th} constraint, $c_j \rightarrow 0$ hardening the constraint. Soft constraints are as follows.

1. *Generic passband/stopband deviation*: For passband frequencies, $H(\omega_{g,j}) - t \leq 1$, and $-H(\omega_{g,j}) - t \leq -1$. For stopband frequencies $H(\omega_{g,j}) - C_s t \leq 0$, and $-H(\omega_{g,j}) - C_s t \leq 0$. The maximum stopband deviation is C_s times that of the passband.
2. *Specific stopband deviation*: Additional suppression is warranted in *other* passband regions, as the latter may contain sinusoids. To this end, for any $\omega_{g,j}$ in another subband filter's passband region, we constrain $H(\omega_{g,j}) - C_{s,0} t \leq 0$, $-H(\omega_{g,j}) - C_{s,0} t \leq 0$, $\partial H(\omega_{g,j})/\partial\omega_{g,j} - C_{s,1} t \leq 0$, and $-\partial H(\omega_{g,j})/\partial\omega_{g,j} - C_{s,1} t \leq 0$. The combination of response and response derivative constraints, given the same length filter, seems to yield better response suppression in these regions than response constraints alone, without detracting from the ability to meet other constraints.

Values $C_s = 2.5$, $C_{s,0} = 0.05$, and $C_{s,1} = 0.25$ produce the example stopband responses shown in Fig. 1.

In Appendix A, we show that optimization with the abovementioned constraints admits a linear programming form. This linear programming optimization, embedded in a semidefinite programming context [11], has complexity which is polynomial time in the filter length. Running on a 1.5 GHz Athlon processor, we find filter lengths over 2000 to be prohibitive.

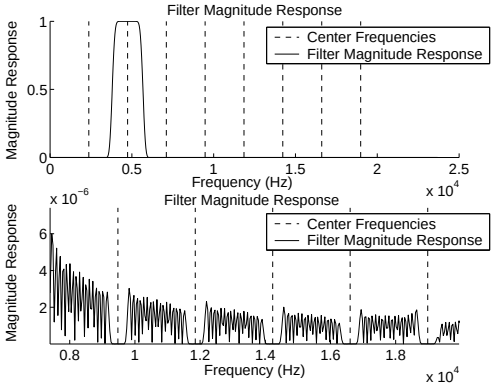

Figure 1: *Optimal filter response and closeup of stopband.*

In Fig. 1, the fundamental frequency is 2375 Hz at a sampling rate of 44100 Hz, and the filter length is 499 samples. However, analyzing signals such as the A0 piano tone (fundamental = 27.5 Hz) generally requires narrower bandwidths and hence longer filters (thousands of samples) where the optimization may be prohibitive. To this end, we pursue a two-stage approach, in which an optimal filter nulls out only the neighboring harmonics, and a secondary bandpass filter of much greater bandwidth provides additional suppression outside the neighboring region. Fig. 2 shows magnitude responses of a prototype nulling filter cascaded with a secondary bandpass filter with -3dB/octave decay.

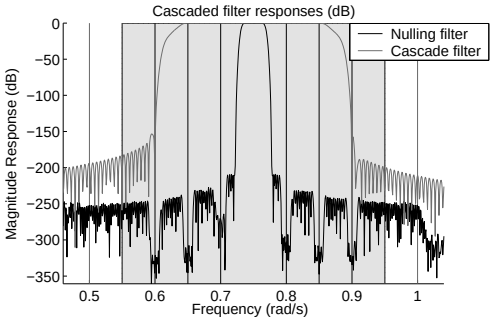

Figure 2: *Nulling filter and secondary bandpass filter.*

In general, we wish to null only the set of center frequencies within a local window of the given partial. A local sliding window contains the same fixed number of center frequencies for all partials plus additional guard bands on both sides, and covers the frequency interval $[\omega_{l,min},\ \omega_{l,max}]$. This range is affinely mapped to as wide a region as possible: $[\omega_{l,min},\ \omega_{l,max}] \rightarrow [\omega'_{min},\ \omega'_{max} = \pi - \epsilon]$ (Fig. 3). Since the nulling filter will be designed in the *primed* space, this ensures the computational cost of the optimization does not become too great.

The nulling filter response is mapped back into the original

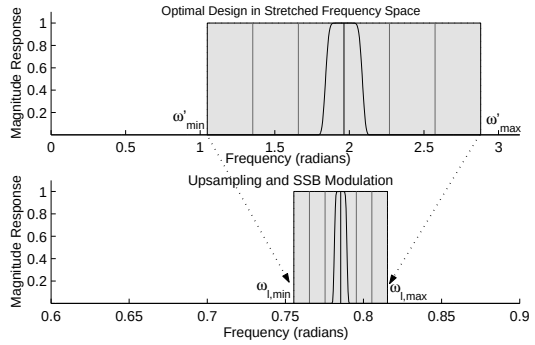

Figure 3: *Mapping of nulling filter responses.*

space by a bandlimited interpolation followed by single sideband (SSB) modulation (Fig. 3). We choose $\omega'_{min} = \pi/3$ by balancing the need for a guard band in the Hilbert transform used in the SSB modulation with the desire to minimize filter length in the primed space.

2.3. Mode frequency and decay estimation

We begin by preprocessing as shown in Fig. 4. The preprocessing goals are twofold: first, to maximize the effective mode frequency separation; second, to undo the coloration of the recording noise by the bandpass filter. Both objectives are met when we extract the analytic signal, heterodyne with respect to the center frequency, and downsample as much as possible. A critically-hopped STFT accomplishes the same objectives, except the subband center frequencies may not relate to the input signal's harmonics.

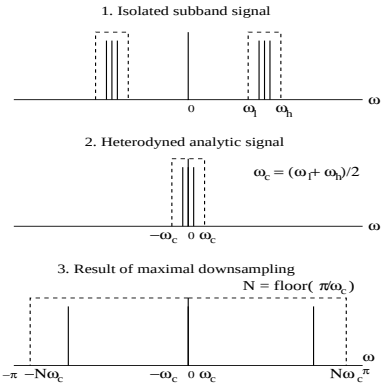

Figure 4: *Preprocessing steps.*

The result of these initial preprocessing steps is an analytic signal which, in the noiseless case, is modeled as the sum of p_l complex, exponentially-decaying sinusoids:

$$y_{l,t} \quad = \quad \sum_{k=1}^{p_l} A_{k,l,t} e^{-\gamma_{k,l} t + j(\omega_{k,l} t + \phi_{k,l,t})}, \; t = 1 \ldots N \quad (5)$$

We estimate $\{\omega_{k,l}\}$ and $\{\gamma_{k,l}\}$ using the Steiglitz-McBride algorithm [9], which fits the subband signal $y_{l,t}$ as the output of a fixed-order IIR recursion driven by an impulse: $\sum_{k=0}^{p_l} b_{k,l} y_{l,t-k} = \sum_{k=0}^{p_l} a_{k,l} \delta_{t-k}$. The poles $\{z_{k,l}\}_{k=1}^{p_l}$ are the roots of $B_l(z) = \sum_k b_{k,l} z^{-k}$. Frequencies and decay factors are extracted from the poles: $\omega_{k,l} = \tan^{-1}(\mathrm{Im}[z_{k,l}]/\mathrm{Re}[z_{k,l}])$, $\gamma_{k,l} = -\log|z_{k,l}|$. The recursion is initialized with poles for which the frequencies equal those of p_l peaks picked from a cepstral-smoothed FFT magnitude spectrum of $y_{l,t}$, and the decay factors equal an estimate, $\hat{\gamma}_l$, of the global decay rate of the subband signal.

To compute $\hat{\gamma}_l$, we first obtain the subband signal's amplitude envelope, $v_{l,t}$, by averaging the minimum and maximum absolute values in length W windows of $y_{l,t}$:

$$v_{l,t} = \frac{\max\left|y_{l,t:\min(N,t+W-1)}\right| + \min\left|y_{l,\max(1,t-W+1):t}\right|}{2} \quad (6)$$

With just one sinusoid, the amplitude envelope is absolute value of the analytic signal: $v_{l,t} = |y_{l,t}|$. However, multiple sinusoids cause ripples in the analytic signal's magnitude envelope, hence the need for the extra envelope detection stage (6).

Note that the subband signals are actually noisy, and that as the sinsoidal components of $y_{l,t}$ decay, they vanish into the noise. Accordingly, our computation of $\hat{\gamma}_l$ only utilizes the first M samples of $v_{l,t}$, where M is an estimate of the index at which the sinusoidal components vanish. We estimate M as the partition point of an optimal piecewise linear fit of $\log v_{l,t}$ according to a weighted least squares criterion; the expectation being that the "signal" part, $\log v_{l,1:M-1}$, will be fit by a line with large negative slope, and the "noise" part, $\log v_{1,M:N}$, will be fit by a line with near zero slope; see Fig. 5. Finally, we set $\hat{\gamma}_l$ to the negative of the slope of the linear fit of $\log v_{l,1:M-1}$.

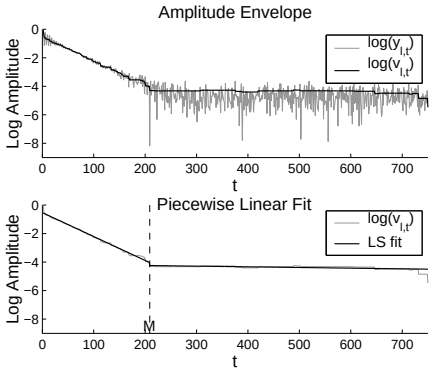

Figure 5: *Amplitude envelope and piecewise linear fit.*

2.4. Amplitude and phase tracking

In order to track amplitudes and phases, we define the following Gaussian state-space model.

$$\begin{aligned} p(s_{l,0}) &\sim \mathcal{N}\left(0,\, \epsilon^{-1}\mathbf{I}\right) \\ p(s_{l,t+1}|s_{l,t}) &\sim \mathcal{N}\left(F_l s_{l,t},\, q\mathbf{I}\right) \\ p(y_{l,t}|s_{l,t}) &\sim \mathcal{N}(H s_{l,t},\, r\mathbf{I}) \end{aligned} \quad (7)$$

Here $y_{l,t} \in \mathbb{R}^{2\times 1}$ represents the real and imaginary parts of the analytic signal ($y_{l,t}(1)$ and $y_{l,t}(2)$, respectively). The state, $s_{l,t} \in \mathbb{R}^{2l_p \times 1}$, represents the component sinusoidal oscillators; here, $s_{l,t}(2k-1)$ represents the real and $s_{l,t}(2k)$ the imaginary part of the k^{th} oscillator. The corresponding amplitudes and phases

are extracted as follows.

$$A_{k,l,t} = \sqrt{s_{l,t}^2(2k-1) + s_{l,t}^2(2k)}$$
$$\phi_{k,l,t} = \tan^{-1}\left[\frac{s_{l,t}(2k)}{s_{l,t}(2k-1)}\right] \quad (8)$$

The sinusoidal dynamics evolve independently for each sinusoid; hence F_l is block diagonal. The k^{th} block is given:

$$F_l(2k{-}1{:}2k, 2k{-}1{:}2k) = e^{-\gamma_{k,l}}\begin{bmatrix}\cos(\omega_{k,l}) & -\sin(\omega_{k,l})\\ \sin(\omega_{k,l}) & \cos(\omega_{k,l})\end{bmatrix} \quad (9)$$

Likewise, $H = [\mathbf{I_2}\ \mathbf{I_2}\ \ldots\ \mathbf{I_2}]$.

In process of estimation, we track $\hat{s}_{l,t} = E(s_{l,t}|y_{l,1:N})$ recursively in time by a Kalman filter and Rauch-Tung-Striebel smoother. The posterior mean $E(s_{l,t}|y_{l,1:N})$ is nothing but the m.m.s.e. estimate of $s_{l,t}$. Substituting $s_{l,t} = \hat{s}_{l,t}$ into (8), obtains amplitude and phase estimates.

Noise parameters $\{q, r\}$ govern the bias-variance tradeoff. A large r/q indicates less variance and more bias, because more observations contribute with similar weights to the estimate $\hat{s}_{l,t}$. As a result, the estimate becomes less sensitive to observation noise, yet it becomes more difficult to track nonstationary amplitudes and phases.

3. CONVERGENCE

A typical convergence profile is shown for a C2 piano tone (fundamental 65 Hz) over five iterations and eight partials, with harmonic initialization and 5% initial fundamental frequency estimation error.

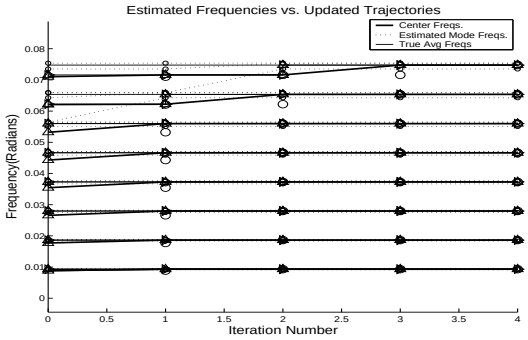

Figure 6: *Filterbank convergence.*

Often, one may want to sacrifice generality in order to improve the convergence rate. If an acoustical model profile is known, e.g. (4), one may estimate its parameters given the collection of average instantaneous frequencies, then simultaneously update *all* filterbank center frequencies according to this profile.

A. LINEARITY OF FREQUENCY RESPONSE CONSTRAINTS

Let $\{h(k)\}$ be the coefficients of the length $(2M{+}1)$ filter; $h(k) = h(-k)$ by the zero-phase constraint. Hence, the free parameters are $x = [h(0) \ldots h(M)\ t]^T$ where t is the tolerance. The objective is to minimize t, such that all constraints of Section 2.2 are satisfied.

In this section, we show the optimization admits a linear programming form; e.g. there exists A, b, c such that, for the optimal x^*: $x^* = \text{argmin}_x\ c^T x$ subject to $Ax \preceq b$, where $\preceq$ means *componentwise* inequality, i.e. $a_j^T x \leq b_j \forall j$.

Since the objective is to minimize t, $c^T = [0\ \ldots\ 0\ 1]$. Now, in Section 2.2 there exist constant b_j, c_j, d_j such that each soft constraint is expressed:

$$d_j \frac{\partial^q H(\omega_{g,j})}{\partial \omega_{g,j}^q} - c_j t \leq b_j \quad (10)$$

for $q \geq 0$. Each hard constraint admits (10) with $c_j = 0$.

It remains to derive $\partial^q H(\omega_{g,j})/\partial\omega_{g,j}^q$ as a linear function of nonnegative filter coefficients $h(0)\ \ldots h(M)$. Since $H(\omega_{g,j}) = \sum_{k=-M}^{M} h(k)e^{-j\omega_{g,j}k} = h(0) + 2\sum_{k=1}^{M}\cos(\omega_{g,j}k)h(k)$, repeated term by term differentiation obtains

$$\frac{\partial^q H(\omega_{g,j})}{\partial \omega_{g,j}^q} = \mathbf{1}_{\{q=0\}}h(0) + 2\sum_{k=1}^{M} k^q T_q(\omega_{g,j}k)h(k) \quad (11)$$

where $T_q(\omega) = \cos(\omega), -\sin(\omega), -\cos(\omega)$ or $\sin(\omega)$ depending on whether $q \mod 4 = 0, 1, 2$ or 3 respectively.

Substituting (11) into (10) expresses each constraint in the form $a_j^T x \leq b_j$, where

$$a_j^T = \begin{bmatrix} d_j 1_{\{q=0\}} & 2d_j T_q(\omega_{g,j}) & 2d_j 2^q T_q(2\omega_{g,j}) \\ & \ldots\ 2d_j M^q T_q(M\omega_{g,j}) & -c_j \end{bmatrix} \quad (12)$$

as was to be shown.

REFERENCES

[1] Aramaki, M., Bensa, J., Daudet, L., Guillemain, Ph., Kronland-Martinet, R., "Resynthesis of coupled piano strings vibrations based on physical modeling", Journal of New Music Research, 30(3): 213-226, 2001.

[2] Bensa, J., Bilbao, S., et al. "From the physics of piano strings to digital waveguides", Proc. 2002 International Computer Music Conference, Göteborg, Sweden, 2002.

[3] Fletcher, H., Blackham, E.D., and Stratton, R. "Quality of piano tones", J. Acoustical Society of America 34(6), 749-761, 1962.

[4] Nakamura, I. "Fundamental theory and computer simulation of the decay characteristics of piano sound", J. Acoustical Society of Japan 10(5), 289-297, 1989.

[5] Portnoff, M. "Implementation of the digital phase vocoder using the fast fourier transform", IEEE Trans ASSP 24(3): 243-248, 1976.

[6] Serra, M.H., Rubine, D., and Dannenberg, R. "Analysis and synthesis of tones by spectral interpolation", J. Audio Eng. Soc., 38(3): 111-127, 1990.

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

[8] Smith, J.O, and Gossett, P. "A flexible sampling-rate conversion method", Proc. ICASSP, 2: 19.4.1-19.4.2, San Diego, March 1984.

[9] Steiglitz, K., and McBride, L.E. "A Technique for the identifi cation of linear systems", IEEE Trans. Automatic Control, AC-10: 461-464, 1965.

[10] Steiglitz, K., Parks, T.W., and Kaiser, J.F. "METEOR: A constraint-based FIR fi lter design program", IEEE Trans. SP 40(8): 1901-1909, 1992.

[11] Vandenberghe L. and Boyd, S. "Positive defi nite programming", SIAM Review, 38(1): 49-95, March 1996.

[12] Weinreich, G. "Coupled piano strings", J. Acoust. Soc. Amer., 62(6), pp. 1474–84, 1977.