MULTI-PITCH ESTIMATION OF AUDIO RECORDINGS USING A CODEBOOK-BASED APPROACH

Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University, Denmark
{mwh,jrj,mgc}@create.aau.dk

ABSTRACT

In this paper, a method for multi-pitch estimation of single-channel mixtures of harmonic signals is presented. Using the method, it is possible to resolve amplitudes of overlapping harmonics, which is otherwise an ill-posed problem. The method is based on the extended invariance principle (EXIP), and a codebook consisting of realistic amplitude vectors. A nonlinear least squares (NLS) cost function is formed based on the observed signal and a parametric model of the signal, for a set of fundamental frequency candidates. For each of these, amplitude estimates are computed. The magnitudes of these estimates are quantized according to a codebook, and an updated cost function is used to estimate the fundamental frequencies of the sources. The performance of the proposed estimator is evaluated using synthetic and real mixtures, and the results show that the proposed method is able to estimate multiple pitches in a mixture of sources with overlapping harmonics.

Index Terms— Multi-pitch estimation, amplitude estimation, vector quantization, music information retrieval.

1. INTRODUCTION

The pitch, or fundamental frequency, is a key feature of harmonic signals, such as short segments of music and speech signals. Music signals often contain multi-pitch signals, e.g., when multiple instruments are playing simultaneously. Pitch estimation has applications in problems such as separation [1], enhancement [2], automatic music transcription [3], and source localization [4].

Two common types of pitch estimation methods exist, i.e., non-parametric methods, and parametric, model-based methods. Examples of single-pitch methods in the former category include methods based on auto-correlation [5, 6]. Auto-correlation based methods have also been used for multi-pitch estimation, an example is [7], which is based on the enhanced summary autocorrelation function (ESACF). However, those methods are sub-optimal from a statistical point of view. Examples of single-pitch methods in the latter category include those based on maximum likelihood (ML) [8, 9] (see [10] for further examples). Parametric multi-pitch methods also exist, and one that uses ML estimation iteratively is the expectation-maximization (EM) algorithm [11], while another is known as the harmonic matching pursuit [12] (see also [10]).

Multipitch estimation becomes difficult when the pitches have overlapping harmonics, for instance in a mixture of two sources where the fundamental frequencies are 300 Hz and 450 Hz. A strong peak would occur at 150 Hz if using, e.g., the NLS estimator, which would result in wrong pitch estimates. A solution might be to map the amplitude estimates to realistic amplitudes in a codebook, e.g., using vector quantization [13]. Vector quantization has previously been applied in parameter estimation of music and speech signals, some notable references include source separation [14], and speech enhancement [15]. Harmonic amplitude information has been used previously in fields such as instrument recognition [16], where the aim is to provide instrument labels for frames with concurrent instruments playing, and automatic music transcription [17, 18], where the aim is to ouput the discrete pitches being played, along with onset times and note durations. Discrete pitch estimates, however, are not useful when estimating the pitch of an instrument played with vibrato, or for the purpose of tuning an instrument.

In this paper, we propose a method for multi-pitch estimation of mixtures of harmonic signals, such as recordings of musical instruments, where harmonics might overlap. In this work, the mixtures are single-channel. The method is based on the extended invariance principle (EXIP) [19, 20], and a codebook of naturally occurring amplitude vectors, trained using amplitude vectors of signals similar to those of interest. The fundamental frequencies are estimated iteratively for each source, and the amplitudes are quantized according to the codebook. The idea is to investigate whether some crude knowledge about the spectral envelope of the components of the mixture signals is beneficial for multi-pitch estimation of musical signals. It should be noted that we are here estimating continuous pitch of the instruments.

The remainder of the paper is organized as follows. In Section 2, the signal model is introduced. The proposed

This work was supported in part by the Villum Foundation, and the Danish Council for Independent Research, grant ID: DFF 1337-00084. This publication only reflects the authors' views.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

multi-pitch estimator is described in Section 3. The experimental setup and results are presented in Section 4, and the work is concluded in Section 5.

2. SIGNAL MODEL

Consider a complex-valued single-channel mixture of M harmonic signals embedded in noise at time instant n. The data can be represented by the snapshot $\mathbf{x} \in \mathbb{C}^N$ i.e.,

$$\mathbf{x} = [x(0) \;\; x(1) \;\; \cdots \;\; x(N-1)]^T. \tag{1}$$

A complex signal model is used because it leads to simpler expressions, and lower computational complexity. It should be noted that although the signal model is complex, it can be used with real signals by applying the Hilbert transform. The entries in the data vector are linear superpositions of M sources, i.e.,

$$x(n) = \sum_{m=1}^{M} s_m(n) + e(n), \tag{2}$$

where

$$s_m(n) = \sum_{l=1}^{L_m} \alpha_{m,l} e^{j\omega_{0,m} l n}, \tag{3}$$

where $\omega_{0,m}$ is the fundamental frequency, L_m the model order (assumed known here, but can be estimated using, e.g., the MAP method, see [10]), and, $l = 1, \ldots, L_m$ is the harmonic index of the mth source, and

$$\alpha_{m,l} = A_{m,l} e^{j\phi_{m,l}} \tag{4}$$

is the complex amplitude, where $A_{m,l}$ is the real amplitude of the lth harmonic for the mth source, $\phi_{m,l}$ its phase, and $e(n)$ is assumed to be white Gaussian noise. It is assumed that the signal is stationary during the interval $n = 0, \ldots, N-1$. A vector signal model can be stated as

$$\mathbf{x} = \sum_{m=1}^{M} \mathbf{Z}_m(\omega_{0,m}) \boldsymbol{\alpha}_m + \mathbf{e}, \tag{5}$$

where $\mathbf{Z}_m(\omega_{0,m})$ is a Vandermonde matrix, i.e.,

$$\mathbf{Z}_m(\omega_{0,m}) = [\mathbf{z}_{m,1}(\omega_{0,m}) \quad \cdots \quad \mathbf{z}_{m,L_m}(\omega_{0,m})], \tag{6}$$

where $\mathbf{z}_{m,l}(\omega_{0,m}) = [1 \;\; e^{j\omega_{0,m} l} \;\; \cdots \;\; e^{j\omega_{0,m} l(N-1)}]^T$. The vector of complex amplitudes is

$$\boldsymbol{\alpha}_m = [\alpha_{m,1} \;\; \cdots \;\; \alpha_{m,L_m}]^T, \tag{7}$$

and

$$\mathbf{e} = [e(0) \;\; e(1) \;\; \cdots \;\; e(N-1)]^T. \tag{8}$$

The likelihood function of the observed signal, parametrized by

$$\boldsymbol{\theta} = [\omega_{0,1} \;\; \boldsymbol{\alpha}_1^T \;\; \cdots \;\; \omega_{0,M} \;\; \boldsymbol{\alpha}_M^T]^T, \tag{9}$$

can be written as

$$p(\mathbf{x}; \boldsymbol{\theta}). \tag{10}$$

Here, we are concerned with estimating the set of fundamental frequencies $\boldsymbol{\omega}_0 = [\omega_{0,1} \;\; \cdots \;\; \omega_{0,M}]^T$.

3. PROPOSED METHOD

We will now derive the proposed multi-pitch estimator. For the signal model at hand, we wish to find the parameters of the multi-pitch mixture, i.e.,

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}). \tag{11}$$

For white Gaussian noise, this can be solved using the NLS method, i.e.,

$$\widehat{\boldsymbol{\omega}}_0 = \arg\min_{\omega_{0,m} \in \Omega} \left\| \mathbf{x} - \sum_{m=1}^{M} \mathbf{Z}_m \boldsymbol{\alpha}_m \right\|_2^2, \tag{12}$$

where Ω is the set of possible frequencies. However, this is a complicated problem to solve for all $\omega_{0,m}$ at once. One possible approach for estimating the parameters is to use an iterative approach, such as the harmonic matching pursuit [12, 10], which we will use. It is based on a residual for iteration i, defined as

$$r^{(i)}(n) = x(n) - \sum_{m=1}^{i} \sum_{l=1}^{L_m} \alpha_{m,l} e^{j\omega_{0,m} l n}, \tag{13}$$

which for $i = 1, \ldots, M$ can be written as

$$r^{(i)}(n) = r^{(i-1)}(n) - \sum_{l=1}^{L_i} \alpha_{i,l} e^{j\omega_{0,i} l n}, \tag{14}$$

and is used to estimate the model parameters iteratively for each source. The method is initialized using the observed signal, i.e., $r^{(0)}(n) = x(n)$. The parameters, for sources $m = 1, \ldots, M$, are then estimated by solving

$$\widehat{\omega}_{0,m} = \arg\min_{\omega_{0,m}, \boldsymbol{\alpha}_m} \left\| \mathbf{r}^{(i-1)} - \mathbf{Z}_m \boldsymbol{\alpha}_m \right\|_2^2, \tag{15}$$

where $\mathbf{r}^{(i)}$ is a vector containing the residual. It should be noted that the cost function is multi-modal, and we therefore perform the minimization using a grid search. The LS estimates of the amplitudes $\boldsymbol{\alpha}_m$ are [21]

$$\widehat{\boldsymbol{\alpha}}_m = \left(\mathbf{Z}_m^H \mathbf{Z}_m \right)^{-1} \mathbf{Z}_m^H \mathbf{r}^{(i-1)}. \tag{16}$$

The estimate of $\omega_{0,m}$ is found by substituting the above into (15), i.e.,

$$\widehat{\omega}_{0,m} = \arg\min_{\omega_{0,m}} \left\| \mathbf{r}^{(i-1)} - \mathbf{Z}_m \left(\mathbf{Z}_m^H \mathbf{Z}_m \right)^{-1} \mathbf{Z}_m^H \mathbf{r}^{(i-1)} \right\|_2^2. \tag{17}$$

The fundamental frequencies and amplitudes of the M sources are then obtained by computing the residual (14) and estimating the fundamental frequency using (17) and the amplitudes using (16). However, estimating the amplitudes

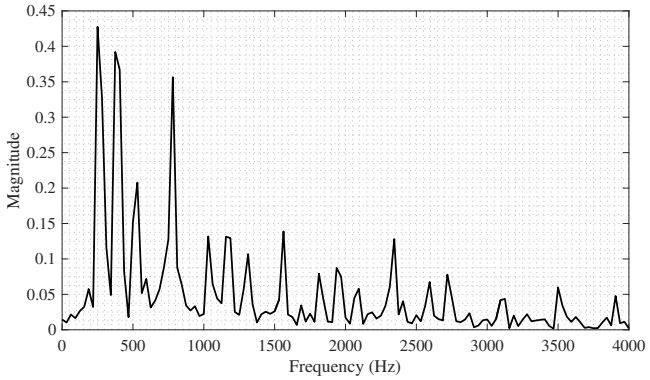

Fig. 1. Spectrum of synthetic signal used for evaluation of the proposed method.

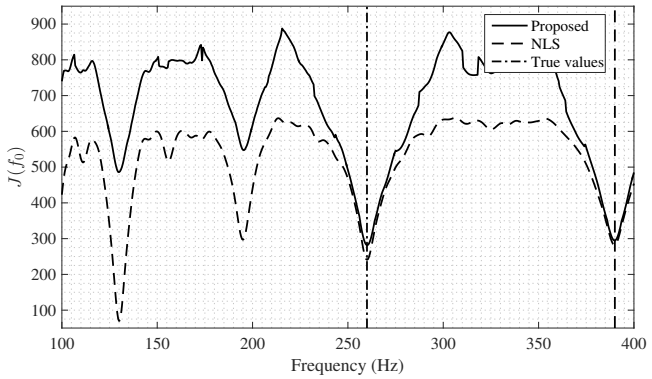

Fig. 2. Initial cost function according to (12) (dotted), where $f_{0,1} = 260$ Hz, and $f_{0,2} = 390$ Hz, and refined cost function according to (21) (solid).

of overlapping harmonics is an ill-posed problem. To solve this, we propose to make use of the EXIP [19, 20], and to map the vector $\widehat{\mathbf{A}}_m$, where each entry is the magnitude of the corresponding entry in $\widehat{\boldsymbol{\alpha}}_m$ to entries in a codebook of realistic amplitudes using a vector quantizer, i.e.,

$$\widehat{\mathbf{A}}_m \rightarrow \widetilde{\mathbf{A}}_m \in \mathcal{C}. \tag{18}$$

In this work, the mapping of amplitudes $\widehat{\boldsymbol{\alpha}}_m$ to codebook entries is done by solving

$$\widetilde{\mathbf{A}}_m = \underset{\widetilde{\mathbf{A}}_m \in \mathcal{C}}{\arg\min} \left\| \widehat{\mathbf{A}}_m - \widetilde{\mathbf{A}}_m \right\|_2^2. \tag{19}$$

It should be noted that the amplitude vectors should be scaled, to limit the size of the codebook. The codebook amplitudes $\widetilde{\mathbf{A}}_m$ are combined with the phases of the amplitude estimates $\widehat{\boldsymbol{\alpha}}_m$ to result in the amplitude estimates

$$\widetilde{\boldsymbol{\alpha}}_m = [\widetilde{A}_{1,m} e^{j\angle\widehat{\alpha}_{1,m}} \quad \cdots \quad \widetilde{A}_{L_m,m} e^{j\angle\widehat{\alpha}_{L_m,m}}]^T. \tag{20}$$

These amplitudes can be substituted in (15), i.e.,

$$\omega_{0,m} = \underset{\omega_{0,m}}{\arg\min} \left\| \mathbf{r}^{(i-1)} - \mathbf{Z}_m \widetilde{\boldsymbol{\alpha}}_m \right\|_2^2. \tag{21}$$

As an example of what we want to avoid, the magnitude of the amplitude of the fundamental frequency should not be allowed to evolve non-smoothly over time. Using the approach proposed here, the magnitudes of the harmonic amplitudes are constrained to have values that would be considered realistic.

The method proposed in this section, which is based on the harmonic matching pursuit [12], could be used to initialize an EM algorithm, where the superimposed signals are harmonic sources [11] (see also [10]).

4. EXPERIMENTS

We now present the experimental evaluation of the proposed multi-pitch estimator. In the initial, proof-of-concept experiment, the data is synthetically generated using the multi-pitch harmonic signal model (2). The synthetic signal contained two sources, i.e., $M = 2$ with fundamental frequencies $f_{0,1} = 260$ Hz and $f_{0,2} = 390$ Hz, respectively, and the number of harmonics for each source was $L_1 = L_2 = 6$. White noise was added to result in an SNR of 20 dB. The spectrum of the signal can be seen in Figure 1.

This setup gives rise to the harmonics, i.e., $f_{0,1} \cdot 3 = f_{0,2} \cdot 2 = 780$ Hz, and $f_{0,1} \cdot 6 = f_{0,2} \cdot 4 = 1560$ Hz. The magnitudes of the amplitudes are decaying, and they are the same for both sources, i.e., $A_{l_m} = 1/l_m$ for $l_m = 1, \ldots, L_m$. The experiments are carried out using segments of length $N = 200$ samples. The codebook contains the true amplitudes and four other realistic amplitude vectors. It is assumed that the number of sources M is known a priori, although the problem of determining the number of sources can be solved using, e.g., a MAP-based method [10]. Figure 2 (dotted line) shows the initial cost function (14) when the input signal is as described above. The data in the figure shows minima at 260 Hz and 390 Hz, but also a very strong minimum at 130 Hz. The amplitudes corresponding to the global minimum at 130 Hz would be $\{0, 1, 1, 1/2, 0, 5/6\}$. However, these are not realistic amplitudes for a real-world signal. By designing the codebook such that none of the codewords have zero (or very small) amplitude for the fundamental frequency of the scenario described, this situation should be avoided. By mapping each vector of initial magnitude amplitude estimates $\widehat{\mathbf{A}}_m$ to the nearest codebook entry (18), this is indeed avoided, as shown in Figure 2 (solid line). The fundamental frequencies estimated using the harmonic matching pursuit [12] using the initial cost function are $f_{0,1} = 130$ Hz, and $f_{0,2} = 260$ Hz, while for the refined cost function the estimates are $f_{0,1} = 260$ Hz, and $f_{0,2} = 390$ Hz. This means that by using the proposed method of mapping magnitude of the initial amplitude estimates to amplitudes in a codebook, we achieve the correct pitch estimates. In a more complex scenario, the results could be used to initialize an expectation-maximization (EM) algorithm [11, 10], which is otherwise not a simple problem.

The proposed method has also been evaluated using real

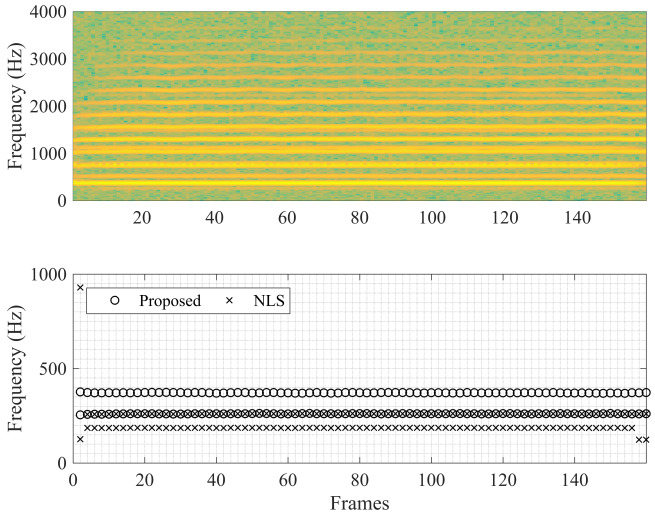

Fig. 3. Spectrogram (top) and pitch estimates (bottom) of a multi-pitch mixture of two instruments, trumpet and horn, playing the notes C4 (262 Hz) and F#4 (370 Hz), respectively.

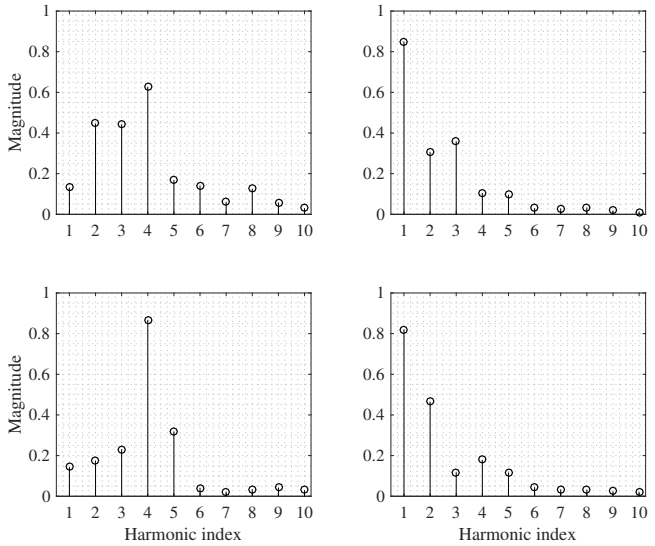

Fig. 4. Four examples of codebook entries, i.e., magnitude amplitudes ($L = 10$).

data[1]. A codebook of amplitudes is trained using ten recordings of different woodwind instruments playing a succession of notes, ranging from C4 (262 Hz) to B4 (494 Hz), i.e., 12 notes. The recordings are single-channel with $f_s = 44.1$ kHz, however, they are downsampled to $f_s = 8$ kHz. The approximate NLS joint pitch and model order estimator in [10] has been used to jointly estimate the pitch and model order for segments of length $N = 240$ samples. The pitch and model order estimates are then used to form LS estimates of the amplitudes (16) for each frame of each signal, resulting in 11544 amplitude vectors. The amplitudes are scaled such that the norm of each amplitude vector equals one before vector quantization. The chosen codeword is then scaled to match the original amplitudes. The codebook has been trained using the K-means clustering algorithm [22], where the first 10 harmonics of the woodwind signals are considered. If the estimated model order is less that 10, the remaining values are set equal to zero for the corresponding codebook entry. Different choices of the number of clusters for the training of the codebooks have been considered, varying from 1 to 100 clusters. Empirically, a suitable number of codewords was found to be 10, which is the number of clusters used in this experiment. Examples of codebook entries are shown in Figure 4. For test data, a multi-pitch mixture was created by mixing two single note recordings of a Bb trumpet (with vibrato) and a French horn, playing the notes C4 (262 Hz), and F#4 (370 Hz), respectively (it should be noted that the training and test data are disjoint). White noise was added to result in an SNR of 20 dB. A spectrogram of the mixture and the multi-pitch estimates obtained using the proposed method are shown in Figure 3. Each pitch estimate is obtained by performing a grid search from 100 Hz to $f_s/4 = 2000$ Hz, with a spacing of 0.5 Hz, and the results are compared to results obtained using the usual NLS cost function. The results on real data are similar to the results on synthetic data. When the amplitudes are estimated using LS, the estimates for one of the notes are half of what they should be. Using the proposed method of mapping the estimated amplitudes to codebook entries, it is possible to correctly estimate the fundamental frequencies in the mixture.

[1]Available at `http://theremin.music.uiowa.edu`.

5. DISCUSSION

In this paper, a method for multi-pitch estimation of mixtures of harmonic signals has been proposed. The method is based on the harmonic matching pursuit [12], where an initial cost function, and amplitude estimates for each candidate fundamental frequency are formed. These initial amplitudes are then mapped to entries in a codebook. The codebook has been trained using recordings of woodwind instruments, while the mixture consists of recordings of brass instruments. The results show that by using the proposed multi-pitch estimator it is possible to estimate the pitches of multiple sources in a mixture of harmonic signals. The results of the estimator could be used to initialize the EM algorithm [11, 10], and the method could be used in automatic music transcription, enhancement, and separation systems. Future work includes investigating the choice of the number of harmonic amplitudes to include in the codebook, e.g., by using a technique such as variable-dimension vector quantization (VDVQ) [23]. Furthermore it should be investigated whether the amplitudes can be modeled statistically instead of using a codebook approach, which involves training, e.g., by using linear prediction [24].

6. REFERENCES

[1] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[2] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Joint filtering scheme for nonstationary noise reduction," in *Proc. European Signal Processing Conf.*, 2012, pp. 2323–2327.

[3] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, New York, 2006.

[4] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 923–933, 2013.

[5] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 1, pp. 24–33, Feb 1977.

[6] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[7] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov 2000.

[8] M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate," in *Proc. Symp. Comput. Process. Commun.* 1969, vol. XIX, pp. pp. 779–797, Polytechnic Press: Brooklyn, New York.

[9] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic cramer-rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 2048–2059, Aug 1997.

[10] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, Synthesis lectures on speech and audio processing. Morgan & Claypool Publishers, 2009.

[11] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the em algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, Apr 1988.

[12] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, Jan 2003.

[13] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[14] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. 5, pp. V–V.

[15] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments.," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 2, pp. 441–452, 2007.

[16] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 116–128, Jan 2008.

[17] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, Nov 2003.

[18] E. Benetos and S. Dixon, "Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1111–1123, Oct 2011.

[19] P. Stoica and T. Söderström, "On reparametrization of loss functions used in estimation and the invariance principle," *Signal Process.*, vol. 17, pp. 383–387, 1989.

[20] M. G. Christensen, "Metrics for vector quantization-based parametric speech enhancement and separation," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3062–3071, 2013.

[21] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, Feb 2000.

[22] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan 1980.

[23] W. C. Chu, "Vector quantization of harmonic magnitudes in speech coding applications - a survey and new technique.," *EURASIP J. on Advances in Signal Processing*, vol. 2004, no. 17, pp. 2601–2613, 2004.

[24] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.