ESTIMATION OF MULTIPLE PITCHES IN STEREOPHONIC MIXTURES 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 stereophonic mixtures of multiple harmonic signals is presented. The method is based on a signal model which takes the amplitude and delay panning parameters of the sources in a stereophonic mixture into account. Furthermore, the method is based on the extended invariance principle (EXIP), and a codebook of realistic amplitude vectors. For each fundamental frequency candidate in each of the sources, the amplitude estimates are mapped to entries in the codebook, and the pitch and model order are estimated jointly. The performance of the proposed method is evaluated using mixtures of real signals. Experiments show an increase in performance when knowledge about the panning parameters is utilized together with the codebook of magnitude amplitudes when compared to a state-of-the-art transcription method.

Index Terms— Multi-pitch estimation, music information retrieval, model-order selection, vector quantization.

1. INTRODUCTION

Often, music signals contain multiple pitches, or fundamental frequencies, e.g., when multiple instruments are played simultaneously. Knowing these fundamental frequencies is useful in many applications where the harmonic signal model is used in, such as, enhancement [1], source localization [2], automatic music transcription [3] and source separation [4].

Several multi-pitch estimation methods exist, e.g., non-parametric methods, such as those based on the autocorrelation function (ACF) estimation, like [5], and statistical parametric, model-based approaches, such as the maximum likelihood (ML) method [6], which can be used iteratively to resolve multiple fundamental frequencies, using, e.g., the harmonic matching pursuit (HMP) [7], and the expectation-maximization (EM) algorithm [6]. Within the area of automatic music transcription, the main goal is to form score-like representations [3], resulting in discrete pitch estimates, even though the pitch is a continuous parameter. Such methods are often based on spectrogram factorization methods, where an input time-frequency representation is decomposed into note templates and activations. Examples are methods based on non-negative matrix factorization (NMF) [8] and probabilistic latent component analysis (PLCA) [9, 10]. Estimating multiple concurrent pitches is a difficult problem, especially when the fundamental frequencies of the sources have overlapping harmonics, or when they are related in a simple way. A method for multi-pitch estimation of recordings of piano signals, where overtones might overlap, is presented in [11]. The method is based on a smooth autoregressive model of the spectral envelope of the overtones of each note. The spectral smoothness principle is presented in [12].

In this paper, we present a method for multi-pitch estimation for stereophonic mixtures of sources consisting of, possibly multiple, harmonic signals, that might have overlapping harmonics. As opposed to the single-channel methods described in the above, mixtures are here assumed to contain several harmonic signals with amplitude and delay panning applied. The method is based on a multi-channel signal model, where the panning parameters are taken into account. The fundamental frequencies are estimated jointly with the model order, iteratively for each source. The least squares (LS) amplitude estimates are then mapped to entries in a codebook trained using amplitude vectors of monophonic signals, and the fundamental frequency and model order of each source are re-estimated using the mapped amplitudes. In this way, the fundamental frequencies of harmonic sources, with overlapping harmonics, can be resolved. In relation to the work presented in [13], where a codebook-based approach for multi-pitch estimation was proposed, the work presented in this paper is based on a stereophonic signal model, introduced in [14], in which a pitch estimator, which takes the amplitude and delay panning parameters into account when estimating the fundamental frequencies of stereophonic mixtures of single-pitch signals, was proposed. Furthermore, in the work presented here, the model order of each harmonic source is estimated jointly with its fundamental frequency. It should be noted that we are here estimating continuous pitch of the signals considered, resulting in a full parameterization of the signals in the mixture. Furthermore, it should be noted that we here consider the panning parameters known, and we consider estimating the parameters a separate problem.

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

2. SIGNAL MODEL

Consider a complex-valued K-channel mixture at time n. The data in the kth channel is represented by a snapshot $\mathbf{x}_k \in \mathbb{C}^N$, i.e.,

$$\mathbf{x}_k = [x_k(0)\ \ x_k(1)\ \ \cdots\ \ x_k(N-1)]^T,$$

for $k = 0, \ldots, K-1$. It should be noted here that a complex signal model is used because it may lead to simpler expressions, and lower computational complexity. It should also be noted that although the signal model is complex, it can be used with real signals by applying the Hilbert transform. We assume that each snapshot is generated by M sources spatially rendered by applying amplitude and delay panning. An example of an amplitude panning law, which could be applied for a stereophonic mix, i.e., $K = 2$, is [15]

$$g_{k,m} = \begin{cases} \cos\theta_m, & \text{for } k = 0. \\ \sin\theta_m, & \text{for } k = 1. \end{cases} \tag{1}$$

where $k \in \{0, 1\}$ is the channel number for a stereophonic mixture, and θ_m is the angle between the pan direction and the left loud speaker ($k = 0$) for the mth source. The aperture of the loud speakers is assumed to be 90° [15], resulting in equal amplitudes for $\theta_m = 45°$, while only one of the channels will be active when $\theta_m = 0°$ or $\theta_m = 90°$. Delays can also be used to enhance the spatial perception [16, 17], where a delay $\tau_{k,m}$ is added to one of the channels of a source. However, this type of panning is less common than amplitude panning. Furthermore, it should be noted that sources might share panning parameters, e.g., when chords are played. The data in channel k is modeled as a linear superposition of M sources, i.e.,

$$x_k(n) = \sum_{m=1}^{M} g_{k,m} s_m(n - \tau_{k,m}) + e_k(n),$$

with

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

where $\omega_{0,m}$ is the fundamental frequency of the mth source, L_m is the model order, and $\alpha_{m,l} = A_{m,l} e^{j\phi_{m,l}}$ is the complex amplitude, where $A_{m,l}$ is the real amplitude of the lth harmonic of the mth source, $\phi_{m,l}$ its phase. The noise $e_k(n)$ is assumed to be white and complex Gaussian, and the signal is assumed to be stationary during the interval $n = 0, \ldots, N-1$. A vector signal model can then be stated as

$$\mathbf{x}_k = \sum_{m=1}^{M} \mathbf{Z}_m \mathbf{G}(k, m) \boldsymbol{\alpha}_m + \mathbf{e}_k, \tag{2}$$

where $\mathbf{Z}_m$ is a Vandermonde matrix, defined as $\mathbf{Z}_m = [\mathbf{z}_{m,1}\ \cdots\ \mathbf{z}_{m,L_m}]$, $\mathbf{z}_{m,l} = [1\ \ e^{j\omega_{0,m} l}\ \ \cdots\ \ e^{j\omega_{0,m} l(N-1)}]^T$, and

$$\mathbf{G}(k, m) = \begin{bmatrix} g_{k,m} e^{-j\omega_{0,m} f_s \tau_{k,m}} \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots g_{k,m} e^{-jL_m \omega_{0,m} f_s \tau_{k,m}} \end{bmatrix},$$

which is defined by the panning parameters (1) and $\tau_{k,m}$. Furthermore, the vector of complex amplitudes is given by $\boldsymbol{\alpha}_m = [\alpha_{m,1}\ \cdots\ \alpha_{m,L_m}]^T$, and $\mathbf{e}_k = [e_k(0)\ \ e_k(1)\ \ \cdots\ \ e_k(N-1)]^T$. The likelihood of the observed signal, parametrized by $\boldsymbol{\psi} = [\omega_{0,1}\ \ g_{k,1}\ \ \tau_{k,1}\ \ \boldsymbol{\alpha}_1^T\ \ \cdots\ \ \omega_{0,M}\ \ g_{k,M}\ \ \tau_{k,M}\ \ \boldsymbol{\alpha}_M^T]^T$, is defined as $p(\mathbf{x}; \boldsymbol{\psi})$, where $\mathbf{x}$ is obtained by stacking $\mathbf{x}_k$, for all channels. Here, we are interested in estimating the set of fundamental frequencies $\boldsymbol{\omega}_0 = [\omega_{0,1}\ \cdots\ \omega_{0,M}]^T$, while the other parameters are considered nuisance parameters.

3. PROPOSED METHOD

We now derive the joint multi-channel multi-pitch and model order estimator. We wish to find the parameters of the multi-pitch mixture, i.e., $\widehat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \ln p(\mathbf{x}; \boldsymbol{\psi})$. For white Gaussian noise, and mutually independent parameters, this resembles the NLS method, i.e.,

$$\widehat{\boldsymbol{\omega}}_0 = \underset{\{\omega_{0,m}\} \in \Omega}{\arg\min} \left\| \mathbf{x} - \sum_{m=1}^{M} \mathbf{Z}_m \mathbf{G}(k, m) \boldsymbol{\alpha}_m \right\|_2^2, \tag{3}$$

where Ω is the set of possible frequencies. However, solving (3) for all $\omega_{0,m}$ at once is a multidimensional problem. One possible approach for estimating the parameters is to use an iterative method, such as the harmonic matching pursuit [7, 6], which the proposed method is based on. The HMP is based on a residual at iteration i at time n, defined as

$$r_k^{(i)}(n) = r_k^{(i-1)}(n) - \sum_{l=1}^{L_i} g_{k,m} \alpha_{m,l} e^{j\omega_{0,m} l (n - \tau_{k,m})}, \tag{4}$$

and is used to estimate the model parameters iteratively for each modeled harmonic source m. The method is initialized using the observed signal, i.e., $r_k^{(0)}(n) = x_k(n)$. As previously mentioned, the fundamental frequencies of the M sources are estimated jointly with the model order. The MAP model selection criterion [18, 6] is used as a model selection rule, i.e.,

$$\widehat{\mathcal{M}}_m = \underset{\mathcal{M}_m}{\arg\min} \sum_{k=0}^{K-1} -\ln p\left(\mathbf{x}_k; \widehat{\psi}_m, \mathcal{M}_m\right) + \frac{1}{2} \ln |\widehat{\mathbf{H}}_m|,$$

where $\widehat{\mathcal{M}}_m$ is the model of the mth source, and $|\cdot|$ denotes the determinant of a matrix. The determinant of the Hessian, $\widehat{\mathbf{H}}_m$, can be approximated using the Fisher information matrix, and a normalization matrix is introduced (see [18]) i.e.,

$$\mathbf{K} = \begin{bmatrix} (N^3 + K^3 - N^2K^2)^{-\frac{1}{2}} & 0 & 0 & 0 \\ 0 & N^{-\frac{1}{2}} & 0 & 0 \\ 0 & 0 & (K^3N)^{-\frac{1}{2}} & 0 \\ 0 & 0 & 0 & N^{-\frac{1}{2}} \mathbf{I}_{2L} \end{bmatrix},$$

where $\mathbf{I}_{2L}$ is a $2L \times 2L$ identity matrix, such that

$$\ln |\widehat{\mathbf{H}}_m| = \ln |\mathbf{K}^{-2}| + \ln |\mathbf{K}\widehat{\mathbf{H}}_m\mathbf{K}|, \quad (5)$$

where the last term, which is of order $\mathcal{O}(1)$, is ignored, and the first term is used as a penalty term. We can now state the joint pitch and model order estimator used to compute initial estimates for sources $m = 1, \ldots, M$, i.e.,

$$\left\{\widehat{\omega}_{0,m}, \widehat{L}_m\right\} = \underset{\boldsymbol{\alpha}_m, \{\omega_{0,m}, L_m\}}{\arg\min} \frac{\ln |\mathbf{K}^{-2}|}{2} + N\sum_{k=0}^{K-1} \ln \left\|\boldsymbol{\beta}_{k,m}\right\|_2^2, \quad (6)$$

where $\boldsymbol{\beta}_{k,m} = \mathbf{r}_k^{(m-1)} - \mathbf{Z}_m\mathbf{G}(k,m)\boldsymbol{\alpha}_m$, and $\mathbf{r}_k^{(m)} = [r_k^m(0)\ r_k^m(1)\ \cdots\ r_k^m(N-1)]^T$. It should be noted that the cost function is multi-modal, and we therefore perform the minimization with respect to $\omega_{0,m}$ using a grid search. The LS estimates of the amplitudes $\boldsymbol{\alpha}_m$ for each candidate $\omega_{0,m}$ are [19]

$$\widehat{\boldsymbol{\alpha}}_m = \left[\sum_{k=0}^{K-1} \mathbf{G}^H(k,m)\mathbf{Z}_m^H\mathbf{Z}_m\mathbf{G}(k,m)\right]^{-1} \cdot \sum_{k=0}^{K-1} \mathbf{G}^H(k,m)\mathbf{Z}_m^H\mathbf{r}_k^{(m-1)}, \quad (7)$$

which are estimated for each of the m sources. The fundamental frequencies and amplitudes of the M sources are then obtained by computing the residual (4) and estimating the fundamental frequency using (6) and the amplitudes using (7). However, estimating the amplitudes of overlapping harmonics is an ill-posed problem. To solve this, we propose mapping 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, each with unit norm, using a vector quantizer, i.e.,

$$\widehat{\mathbf{A}}_m \rightarrow \overline{\mathbf{A}}_m \in \mathcal{C}.$$

In principle this can be done for all possible $\omega_{0,m}$, but to reduce the computational requirements, we restrict the possible fundamental frequency candidates to be the 100 minima of the cost function in (6). In this work, the mapping of amplitudes $\widehat{\boldsymbol{\alpha}}_m$ to codebook entries is done, according to the EXIP [20, 21], by finding

$$\widetilde{\mathbf{A}}_m = \min_{\gamma_m \in \mathbb{R}^+, \overline{\mathbf{A}}_m \in \mathcal{C}} \left\|\widehat{\mathbf{A}}_m - \gamma_m\overline{\mathbf{A}}_m\right\|_2^2, \quad (8)$$

where γ_m is a scaling factor, to limit the size of the codebook. The codebook is generated by jointly estimating the fundamental frequency and the model order of a set of recordings of monophonic signals. The dimension of the amplitude vectors varies with the model order and the fundamental frequency. To further limit the size of the codebook, the dimension of the amplitude vectors is converted to fixed dimension using a non-square transform, in this case zero padding, if the model order is less than the fixed dimension, and truncation vice versa [22]. The amplitudes $\widetilde{\mathbf{A}}_m$ in (8) are combined with the phases of the initial amplitude estimates $\widehat{\boldsymbol{\alpha}}_m$ to result in the amplitude estimates of the mth source, i.e.,

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

These amplitudes can be substituted in (6), to obtain refined estimates of the fundamental frequency and model order of source m, i.e.,

$$\left\{\widetilde{\omega}_{0,m}, \widetilde{L}_m\right\} = \underset{\boldsymbol{\alpha}_m, \{\omega_{0,m}, L_m\}}{\arg\min} \frac{\ln |\mathbf{K}^{-2}|}{2} + N\sum_{k=0}^{K-1} \ln \left\|\widetilde{\boldsymbol{\beta}}_{k,m}\right\|_2^2 \quad (9)$$

where $\widetilde{\boldsymbol{\beta}}_{k,m} = \mathbf{r}_k^{(m-1)} - \mathbf{Z}_m\mathbf{G}(k,m)\widetilde{\boldsymbol{\alpha}}_m$. As an example of what we want to avoid, the magnitude of the amplitudes of the harmonics should not be allowed to evolve non-smoothly across frequencies, i.e., the spectral smoothness principle is used [12]. 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 a modification of the harmonic matching pursuit [7], could be used to initialize an EM algorithm, to yield better estimates [6]. It should be noted that the panning parameters are assumed known in this work, however, a method for joint DOA and pitch estimation, such as the one in [2] could be used to estimate the panning parameters. Furthermore, it could be exploited that the parameters evolve slowly over time, to allow processing of larger chunks of the signals.

4. EXPERIMENTS

We now present the experimental setup along with the evaluation of the proposed multi-pitch estimator. The experiments have been conducted using mixtures of real recordings of a Bb trumpet (played with vibrato) and a French horn, from the IOWA database[1]. Data from the MAPS database of piano signals [11] has also been used for the evaluation. The mixtures generated using the IOWA database each contain recordings of four notes played simultaneously ($M = 4$). A codebook of amplitudes is trained using 10 recordings of different woodwind instruments each playing a succession of notes, ranging from C4 (262 Hz) to B4 (494 Hz). The recordings are single-channel with $f_s = 44.1$ kHz, however, they are downsampled to $f_s = 8$ kHz. The ANLS joint pitch and model order estimator in [6] 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 (7) for each frame of each signal, resulting in 11544 amplitude vectors. Each amplitude vector is scaled to have unit norm before vector quantization. The chosen codeword is then scaled to match the original amplitude vector. The codebook has been trained using K-means

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

Table 1. GERs for the experiment with IOWA mixtures. The chords are listed using integer notation (t: trumpet, h: horn).

Notes (t-t-h-h)	Proposed	BW2015	ESACF
0-4-7-11	0.0207	0.4861	0.4345
0-3-7-8	0.0143	0.4226	0.4741
0-4-5-9	0.0143	0.5000	0.2378
0-1-5-8	0.0256	0.4980	0.2363
0-3-7-10	0.0156	0.5119	0.6692
0-4-7-9	0.0255	0.3393	0.2378
0-3-5-8	0.0321	0.4782	0.2241
0-2-5-9	0.0223	0.3254	0.4726

[23], where 15 harmonics of the woodwind signals are considered. The dimension of the codebook is converted using a non-square transform, as described in the previous section. 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 20, which is the number of clusters used here.

Two experiments were conducted. In the first experiment, signals are generated by mixing two recordings of notes played using a Bb trumpet, and two notes played using a French horn, i.e., four notes, together forming a 7th chord, using data from the IOWA database. It should be noted that the training data set used for training the codebook, and the test data set used to generate the mixtures are disjoint. In total, eight mixtures are generated. The choice of notes is done in a way similar to in [11]. Figure 1 shows an example of a spectrogram of such a mixture, with pitch estimates. Stereo versions of the mixtures of trumpet and horn signals were generated by applying amplitude and delay panning, as described in Section 2. As mentioned, the panning parameters are assumed known in this work, however, the parameters can be found by adding search dimensions to (9). The panning parameters of the trumpet submixture are $\theta = 25°$, $\tau_0 = 0$ ms, $\tau_1 = 18$ ms while for the horn submixture, they are $\theta = 65°$, $\tau_0 = 18$ ms, $\tau_1 = 0$ ms. For the proposed method, the fundamental frequencies are obtained by performing a grid search from 100 Hz to $fs/4 = 2000$ Hz, with a step size of 1 Hz. The performance of the proposed method has been compared to the method presented in [10], which we will denote BW2015[2] in the figures, and the MIRtoolbox [24] implementation of the ESACF method [5]. For each mixture, the gross error rate (GER) is calculated, which is the number of fundamental frequencies that deviate more than a semitone relative to the ground truth, which is generated using the joint ANLS estimator [6]. The results are shown in Table 1.

In the second experiment, data from the MAPS database,

[2]The source code is available at `https://code.soundsoftware.ac.uk/projects/amt_plca_5d`.

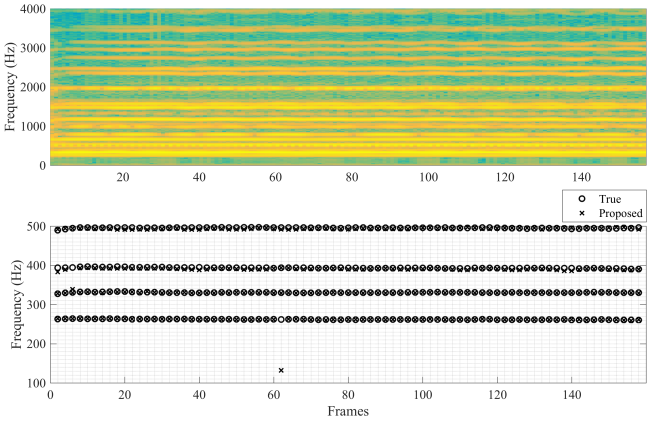

Fig. 1. Spectrogram (top) and pitch estimates (bottom) of a multi-pitch mixture of two instruments, trumpet and horn, playing the notes C4 (262 Hz), E4 (330 Hz), G4 (392 Hz) and B4 (494 Hz), respectively.

i.e., recordings of a set of two-note chords using the *ENST-DkCl* piano, was used. Eight recordings containing signals with fundamental frequencies ranging from C3 (131 Hz) and B5 (988 Hz) were chosen. The data used is single-channel, i.e, $K = 1$, and the number of sources is $M = 2$. This type of evaluation is similar to the one in [13], however, here the pitch is estimated jointly with the model order. The metric used is similar to the one used in the first experiment. The mean GERs were 0.4064 for the proposed method, and 0.2006, for the method presented in [10], respectively.

5. DISCUSSION

In this paper, a method for joint multi-pitch and model order estimation of delay and amplitude panned mixtures of harmonic sources has been proposed. The method presented here extends the work in [14] and [13], where stereophonic mixtures of monophonic sources, and single-channel multi-pitch mixtures were considered, respectively. The proposed method is based on a signal model that takes the panning parameters of a mixture into account. Furthermore, a codebook of amplitude vectors is used to quantize the magnitude of the amplitudes when estimating the multiple fundamental frequencies. For the IOWA mixtures considered, the proposed method outperforms the methods to which it has been compared to, with mean GERs of 0.0331 for the proposed method, 0.4452 for the BW2015 method [10], and 0.3733 for the ESACF method [5], respectively. In the second experiment, with piano data, the BW2015 method outperforms the proposed method. However, it should be noted that the proposed method is based on a harmonic signal model, whereas piano signals can be considered to be quite inharmonic. Furthermore, since the proposed method estimates continuous pitch, it is possible to observe tonal details, such as vibrato.

6. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

[8] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *in Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[9] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a convolutive probabilistic model," in *8th Sound and Music Computing Conference*, 2011, pp. 19–24.

[10] E. Benetos and T. Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *16th International Society for Music Information Retrieval Conference (ISMIR)*, October 2015, pp. 701–707.

[11] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1643–1654, Aug 2010.

[12] 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.

[13] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Multi-pitch estimation of audio recordings using a codebook-based approach," in *Proc. European Signal Processing Conf.*, 2016.

[14] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Pitch estimation of stereophonic mixtures of delay and amplitude panned signals," in *Proc. European Signal Processing Conf.*, 2015.

[15] V. Pulkki, *Spatial sound generation and perception by amplitude panning techniques (PhD thesis)*, Helsinki University of Technology, 2001.

[16] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[17] B. Katz, *Mastering Audio - The Art and the Science*, Focal Press, 2007.

[18] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[19] 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.

[20] 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.

[21] 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.

[22] C. Li, P. Lupini, E. Shlomot, and V. Cuperman, "Coding of variable dimension speech spectral vectors using weighted nonsquare transform vector quantization," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 622–631, Sep 2001.

[23] 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.

[24] O. Lartillot and P. Toiviainen, "A MATLAB toolbox for musical feature extraction from audio," in *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, 2007.