Estimation of Fundamental Frequencies in Stereophonic Music Mixtures

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

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2018.2878384

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Estimation of Fundamental Frequencies in Stereophonic Music Mixtures

Martin Weiss Hansen, *Student Member, IEEE,* Jesper Rindom Jensen, *Member, IEEE,* and Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*—In this paper, a method for multi-pitch estimation of stereophonic mixtures of harmonic signals, e.g., instrument recordings, is presented. The proposed method is based on a signal model which includes the panning parameters of the sources in a stereophonic mixture, such as those applied artificially in a recording studio. If the sources in a mixture have different panning parameters, this diversity can be used to simplify the pitch estimation problem. The mixing parameters of the sources might be shared, resulting in a multi-pitch estimation problem, which is solved using an approach based on an expectation-maximization algorithm for Gaussian sources, where the fundamental frequencies and model orders are estimated jointly. The fundamental frequencies may be related, resulting in overlapping harmonics, complicating the estimation of the parameters. A codebook of harmonic amplitude vectors is trained on recordings of instruments playing single notes, and used when estimating the amplitudes of the mixture components. The proposed method is evaluated using stereophonic mixtures of instrument recordings, and is compared to state-of-the-art transcription and multi-pitch estimation methods. Experiments show an increase in performance when knowledge about the panning parameters is taken into account. The proposed method provides a full parametrization of the components of the observed signal. Possible applications include instrument tuning, audio editing tools, modification of harmonic mixture components, and audio effects.**

***Index Terms*—Multi-pitch estimation, multi-channel pitch estimation, music information retrieval, model selection, vector quantization, sterephonic signal analysis.**

I. Introduction

THE fundamental frequency, or pitch, of a periodic signal, e.g., a short segment of recorded speech or music, is related to the period with which the signal repeats itself. Commonly occurring signals often contain multiple such signals, e.g., recordings of multiple speakers that talk simultaneously, or music recordings where several instruments are active at the same time, which complicates the estimation of the fundamental frequencies of the sources. Determination of the fundamental frequencies of the individual harmonic sources facilitates many tasks, e.g., automatic music transcription [1], source separation [2]–[4], classification of music [5], instrument recognition [6], enhancement [7], and localization [8].

The main types of methods for (single-channel) pitch estimation are non-parametric methods, parametric methods, and methods based on the human auditory system. Examples of non-parametric methods for pitch estimation include those based on auto-correlation [9], and cross-correlation [10]. Such methods are generally prone to octave errors, since they compare a signal to a delayed version of the same signal, or a modified version of the signal. Parametric methods, as the name suggests, are based on parametric signal models. Several types of parametric methods exist, e.g., stastistical methods, like those based on maximum likelihood (ML), see, e.g., [11], and Bayesian methods. Other parametric methods are based on filtering approaches, such as comb filtering, and optimal filtering. A third class of parametric methods is based on subspace methods, examples are Multiple Signal Classification (MUSIC) and Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT), which are based on decomposing the observed signal using subspace approaches, see, e.g., [11] and the references herein. Another class of methods are based on the human auditory system. An example is the approach proposed in [12].

In terms of multi-pitch estimation, some notable existing methods are non-parametric methods, such as those based on the autocorrelation function (ACF), see, e.g., [13], and statistical, parametric approaches, such as the maximum likelihood (ML) method [11], which can be used iteratively to resolve multiple fundamental frequencies, using, e.g., the harmonic matching pursuit (HMP) [14], and the expectation-maximization (EM) algorithm [11]. Within the area of automatic music transcription, the main goal is to form score-like representations [1], 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) [15] and probabilistic latent component analysis (PLCA) [16], [17].

Estimating multiple concurrent pitches is a difficult problem, especially when the sources share energy at some of their harmonics, i.e., when the fundamental frequencies are related in a simple way, as is often the case for music signals. A method for multi-pitch estimation of recordings of piano signals, where overtones might overlap, is presented in [18]. The method is based on a smooth autoregressive model of the spectral envelope of the harmonics of each note. The spectral smoothness principle is presented in [19]. An NMF-based method for multi-pitch estimation, which takes spectral smoothness into account is presented in [20]. Another approach is presented in [21], where a structure (block sparsity) is imposed on the components in a multi-pitch signal. The

M. W. Hansen, J. R. Jensen, and M. G. Christensen are with the Audio Analysis Laboratory, Department of Architecture, Design and Media Technology, Aalborg University, Aalborg DK-9000, Denmark (e-mail mwh@create.aau.dk; jrj@create.aau.dk; mgc@create.aau.dk).

Manuscript received X, X; revised X, X.

fundamental frequencies are found by solving an optimization problem. In [22], a framework is presented which allows incorporating prior information in source separation. However, to estimate the fundamental frequencies of the sources an additional step is required where the output of the separation algorithm is fed to a pitch estimation algorithm.

Most of the music recordings available are recorded in stereophonic format, i.e., two channels. Because the signals in the channels typically share information, it makes sense to exploit all available channels of data, and the performance is expected to increase, when taking both channels into account. Within the area of array processing, multiple channels of data are used, e.g., to perform enhancement (see, e.g., [23] and the references herein). Most pitch estimators operate on single channel data, with a few exceptions, including a method based on a non-parametric multi-microphone periodicity function (MPF)[24]. Another is the multi-channel maximum likelihood (MC ML) pitch estimator presented in [25], which allows for different conditions in the channels, thereby increasing the performance. Previously, spatial diversity has been exploited, e.g. in [8], where joint estimation of the direction-of-arrival (DOA) and pitch is considered.

To the authors' knowledge no method has previously been proposed, which exploits the panning parameters of the sources in a stereophonic mixture when estimating the fundamental frequencies of multiple concurrent sources. A stereophonic mixture is typically created in a recording studio by mixing several stereophonic signals, which may contain multiple fundamental frequencies, e.g., when a chord is played on an instrument. Each signal might have different mixing parameters, such as panning parameters and equalization. In this paper, we assume that mixtures are composed of signals that are spatially enhanced by amplitude and/or delay panning. Amplitude panning is a frequently used virtual source positioning technique, where different gains are applied to the individual channels of a signal. The perception of direction is dependent on these gain factors [26]. Furthermore, a delay can be applied to one of the channels of a source to enhance its spatial quality and to add depth [27]. We refer to this effect as delay panning. If a signal is delayed by more than 1 ms in a stereo setup, the perceived direction of the source is determined mostly by the signal which arrives first [28]. According to [27], the spatial quality of a signal is enhanced by using delays in the 12 to 40 ms range. The effect is called the Haas effect [29]. The idea of separating sources from a multi-channel mixture is used within the source separation [30] and array processing [23] research communities but it has, to the knowledge of the authors, not been applied within the area of pitch estimation and its application in, for example, music transcription. Exploiting knowledge regarding the above-mentioned panning parameters should result in more accurate estimates, and increase the performance in complicated scenarios, where multiple sources are active at once. Several sources might share panning parameters, and estimating the fundamental frequencies of such submixtures becomes difficult, especially when the relationship between the fundamental frequencies of multiple sources results in harmonic overlap. For such signals, two sources might be modelled as a single source, using a model more complex than the model of each of the individual sources. A solution might be to include prior knowledge about the amplitude vectors, and to map the amplitude estimates to realistic amplitudes in a codebook, e.g., using vector quantization [31]. Vector quantization has previously been applied in parameter estimation of music and speech signals. Some notable references include source separation [32], and speech enhancement [33]. Harmonic amplitude information has been used previously in fields such as instrument recognition [34], where the aim is to provide instrument labels for frames with concurrent instruments playing, and automatic music transcription [19], [35], where the aim is to output the discrete pitches being played, along with onset times and note durations. In some cases, however, discrete pitch estimates are not sufficient, e.g., if we wish to estimate the pitch of an instrument played with vibrato, which is a slight variation in pitch throughout a note. Discrete pitch estimates are also not useful if the goal is to use a system for instrument tuning purposes.

In this paper, we propose a parametric method for multi-pitch estimation of stereophonic mixtures of sources consisting of, possibly multiple, harmonic signals, where the harmonics of the signals might be related in a simple way, e.g., when the signals share energy at their harmonics. As opposed to the single-channel methods described above, mixtures are here assumed to contain several harmonic signals with amplitude and delay panning applied, such as in studio recordings. The proposed method is based on a multi-channel signal model, where the panning parameters are taken into account. It should be noted that in the method proposed here, the panning parameters are assumed known, as the goal is to investigate how such knowledge can be exploited when estimating multiple fundamental frequencies. The panning parameters can be estimated, e.g., by employing a method such as the one presented in [36], which we use herein. Furthermore, the term delay panning that we use here, covers delays added to the signals in a more general sense, i.e., the delays might not be applied on purpose, but could for instance arise, when recording a band or an orchestra using multiple microphones. The fundamental frequencies and model orders of the sources are estimated iteratively. The least squares (LS) amplitude estimates [37] are 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. This paper extends our previous work, presented in [38], where a method for stereophonic multi-pitch estimation is presented. The paper is related to the work presented in [39], where a codebook-based approach for multi-pitch estimation of single-channel sources was proposed. The work is based on a stereophonic signal model, introduced in [40], in which a pitch estimator, that takes the amplitude and delay panning parameters into account when estimating the fundamental frequencies of stereophonic mixtures of single-pitch signals, was proposed. An application of the proposed method for source separation and re-panning is presented in [41]. The extension to the previously mentioned work includes

the addition of a refinement step by means of an EM algorithm, and a scheme for detecting the number of sources present in the observed signal. Furthmore, the evaluation of the proposed method is extended. It should be stressed that we are here estimating continuous pitch of the signals considered, resulting in a full parameterization of the signals in the mixture. In this work, we assume that a music mixture has been artificially generated, e.g., in a recording studio, by applying amplitude and delay panning to the sources, and we consider estimating the panning parameters a separate problem. A reason for doing so is to allow processing of longer segments of the mixtures to exploit the stationary nature of the panning parameters, which is also shown in [36]. The main objective of this paper is to investigate how exploiting knowledge about the panning parameters influences the performance of the proposed multi-pitch estimator.

The rest of this paper is organized as follows. In Section II, the multi-channel signal model is described, along with the mixing assumptions (artificial studio recordings). In Section III, the proposed multi-channel multi-pitch estimator is presented, along with details on the harmonic amplitude codebook approach, and the detection framework. The experimental validation of the proposed method is presented in Section IV. Finally, Section V concludes the work presented in this paper.

II. 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 channel $k = 1, \ldots, K$. It should be noted here that a complex signal model is used because it may lead to simpler expressions, and a 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 using amplitude and delay panning. An example of an amplitude panning law, which could used to calculate the gains applied to each channel of a stereophonic mixture is [42]

$$g_{k,m} = \begin{cases} \cos\,\theta_m, & \text{for } k = 1. \\ \sin\,\theta_m, & \text{for } k = 2. \end{cases} \tag{1}$$

where $k \in \{1, 2\}$ is the channel index, and θ_m is the angle between the pan direction and the left loud speaker ($k = 1$) for the mth source. The aperture of the loud speakers is here assumed to be 90° [42], with equal gains applied to the channels when $\theta_m = 45°$, while only one of the channels will be active when $\theta_m = 0°$ or $\theta_m = 90°$. As mentioned in Section I, delays can also be used to enhance the spatial perception in several ways [27], [28], by applying a delay of $\tau_{k,m}$ (in seconds) to one of the channels of a source. We call this effect delay panning, even though the delay might also result from other effects being applied to a signal, or from recording instruments in a live setting. The use of delays as a panning effect is less common than amplitude panning. Furthermore, it should be noted that sources might share panning parameters, e.g., when chords are played on an instrument, or when a submixture is created with a group of instruments playing together. In light of this, we define a source as a single harmonic component. A submixture is defined as one or more sources that share panning parameters, and a mixture consists of one or more submixtures. The signal in channel k of the mixture is modelled as a linear superposition of M harmonic sources with amplitude and delay panning applied as described above, i.e.,

$$x_k(n) = \sum_{m=1}^{M} g_{k,m} s_m(n - f_s \tau_{k,m}) + e_k(n), \tag{2}$$

where $g_{k,m}$ and $\tau_{k,m}$ are the panning parameters of the kth channel of the mth source, f_s is the sampling frequency, and $e_k(n)$ contains a noise component. The mth source s_m is modelled as a sum of L_m harmonic components, i.e,

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

where $\omega_{0,m}$ is the fundamental frequency of the mth source, L_m is the model order, and $a_{m,l} = A_{m,l} e^{j\phi_{m,l}}$ is the complex amplitude, where $A_{m,l}$ is the real amplitude of the lth harmonic of the mth source, and $\phi_{m,l}$ its phase. It should be noted that following this definition of a source, a recording of a single instrument can contain multiple sources, such as a chord played on a guitar, where the signal originating from each string is a source according to our definition. The signal model in (3) is harmonic, i.e., with integer harmonic relationship. Some signals, e.g., recordings of string instruments (guitar, violin, piano, etc.), exhibit inharmonicity, and the signal model might not hold exactly. If an inharmonic signal is modelled using a harmonic model, the energy measured at the harmonic frequencies may be slightly less than the energy exactly at the harmonics. For small inharmonicity coefficients, the difference may be small. However, since the instrumentation in an observed mixture is often unknown a priori, it may be reasonable to choose the harmonic signal model over the inharmonic model. See, e.g., [11], [43] for a more thorough discussion of how to incorporate inharmonicity in the signal model. Furthermore, while the focus in this paper is on stereophonic music mixtures, it should be noted that the model used for $g_{k,m}$ can be modified to allow arbitrary relationships between multiple source channels. In terms of effects added to the sources before and/or after mixing, we do not consider this here. For an effect such as mild equalization, we do not expect the performance to suffer in general. However, nonlinear effects such as distortion and dynamic range compression may introduce frequency components not present in the original signals, and this could have a negative impact on the performance, since the method proposed here is based on the harmonic signal model. Furthermore, if reverberation is applied to one or more of the signals in the stereophonic mixture, then the resulting smearing of the spectrogram may in some cases have a negative impact on performance. We note that it may be possible to remove some of the audio effects, see, e.g., [44] regarding the topic of dereverberation. In [45] a method is presented to revert the effect of dynamic range

compression. Since we here focus on modelling the periodic components of the observed signal, this implies that the noise component may contain non-periodicities not accounted for by the signal model. Still, the noise component $e_k(n)$ is assumed to be white and complex Gaussian, because specifying a model which is suitable in general is difficult, and because it is the distribution that maximizes the entropy [46]. Furthermore, the signal is assumed to be stationary during the interval $n = 0, \ldots, N-1$. The signal model can be written in vector form as

$$\mathbf{x}_k = \sum_{m=1}^{M} \mathbf{Z}_m \mathbf{G}_{k,m} \mathbf{a}_m + \mathbf{e}_k, \tag{4}$$

where $\mathbf{Z}_m$ is a Vandermonde matrix with the harmonic components of the source with fundamental frequency $\omega_{0,m}$ in the columns, i.e.,

$$\mathbf{Z}_m = \begin{bmatrix} 1 & \cdots & 1 \\ e^{j\omega_{0,m}} & \cdots & e^{j\omega_{0,m}L_m} \\ \vdots & \ddots & \vdots \\ e^{j\omega_{0,m}(N-1)} & \cdots & e^{j\omega_{0,m}L_m(N-1)} \end{bmatrix},$$

and $\mathbf{G}_{k,m}$ is a diagonal matrix containing the panning parameters in (1) and $\tau_{k,m}$ for channel k of source m, i.e.,

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

It can be seen that when only amplitude panning is applied to a submixture, $\tau_{k,m} = 0 \ \forall \ \{k,m\}$, and when only delay panning is used, $g_{k,m} = 1 \ \forall \ \{k,m\}$. Also, we assume that the panning parameters are constant throughout a segment of the observed mixture. The vector of complex amplitudes is given by

$$\mathbf{a}_m = [a_{m,1} \ \cdots \ a_{m,L_m}]^T, \tag{5}$$

and the noise vector for channel k is

$$\mathbf{e}_k = [e_k(0) \ e_k(1) \ \cdots \ e_k(N-1)]^T. \tag{6}$$

We now derive the log-likelihood of the kth channel of an observed signal parametrized by $\boldsymbol{\psi}_k = [\boldsymbol{\psi}_{k,1} \ \cdots \ \boldsymbol{\psi}_{k,M}]^T$, where $\boldsymbol{\psi}_{k,m} = [\omega_{0,m} \ g_{k,m} \ \tau_{k,m} \ \mathbf{a}_m^T]^T$, for $m = 1, \ldots, M$. We assume that the deterministic part of the signal is stationary, and that the noise is independent and identically distributed over n and k. Furthermore, we assume that the noise is white Gaussian with possibly different variance in each channel, σ_k^2. The likelihood of the kth channel of the observed signal, is defined as

$$p(\mathbf{x}_k; \boldsymbol{\psi}_k) = \frac{1}{(\pi\sigma_k^2)^N} e^{-\frac{1}{\sigma_k^2}\|\mathbf{e}_k\|_2^2}, \tag{7}$$

which across channels becomes

$$p(\{\mathbf{x}_k\}; \{\boldsymbol{\psi}_k\}) = \prod_{k=1}^{K} \frac{1}{(\pi\sigma_k^2)^N} e^{-\frac{1}{\sigma_k^2}\|\mathbf{e}_k\|_2^2}. \tag{8}$$

The log-likelihood of a single channel of the observed signal is

$$\ln p(\mathbf{x}_k; \boldsymbol{\psi}_k) = -N\ln\pi - N\ln\sigma_k^2 - \frac{\|\mathbf{e}_k\|_2^2}{\sigma_k^2} \tag{9}$$

while the log-likelihood for all channels of the observed signal is

$$\ln p(\{\mathbf{x}_k\}; \{\boldsymbol{\psi}_k\}) = -KN\ln\pi - N\sum_{k=1}^{K}\ln\sigma_k^2 - \sum_{k=1}^{K}\frac{\|\mathbf{e}_k\|_2^2}{\sigma_k^2}. \tag{10}$$

The fundamental frequencies, the complex amplitudes, and the noise variance for each channel are estimated by maximizing (10). Although the focus in this paper is on analysis of stereophonic mixtures, it should be noted that the signal model in (2) and (4), and the likelihoods above are applicable to scenarios with arbitrary numbers of channels (the panning law must be chosen accordingly).

III. Proposed Method

A. Overview

The purpose of the proposed method is to estimate the fundamental frequencies of multiple sources that have been combined into a stereophonic mixture, by altering their spatial qualities, i.e., applying gains and delays to the sources. The stereophonic signal model described in the previous section allows taking into account the panning parameters used when creating the mixture. The panning parameters are estimating using a recently proposed method [36]. In Section III-B, an expectation-maximization (EM) algorithm for estimating multiple fundamental frequencies in a segment of a stereophonic mixture is presented. In Section III-C some interpretation of the method along with a fast way of evaluating the cost function is presented. Section III-D presents the method used to initialize the EM algorithm, which is based on harmonic matching pursuit (HMP). In music mixtures, the fundamental frequencies of the sources are often related in a way such that energy will be shared among harmonics of multiple sources. This typically results in erroneous fundamental frequency estimates, since it may happen that multiple sources are modelled by a more complex model (with a lower fundamental frequency). In Section III-E the proposed method for dealing with this issue is presented. The method is based on a codebook of amplitude vectors trained using recordings of single notes played using a variety of instruments. In Section III-F a detection algorithm is presented for estimating the number of harmonic sources in a stereophonic mixture.

B. Stereophonic Multi-Pitch Estimation

Based on the signal model presented in the previous section, we derive the joint multi-channel multi-pitch and model order estimator. We wish to estimate the fundamental frequency $\omega_{0,m}$ and the vector of amplitudes $\mathbf{a}_m$ for each source in the mixture, and the noise variance σ_k^2 in each channel, by maximizing the log-likelihood in (10). Since the noise variance is assumed independent across channels, it can be estimated from each channel by differentiating (9) w.r.t. σ_k^2 and equating with zero, i.e.,

$$\widehat{\sigma}_k^2 = \frac{\|\mathbf{e}_k\|_2^2}{N} = \frac{1}{N}\left\|\mathbf{x}_k - \sum_{m=1}^{M}\mathbf{Z}_m\mathbf{G}_{k,m}\mathbf{a}_m\right\|_2^2, \tag{11}$$

which, in inserted into (10), results in

$$\ln p\left(\{\mathbf{x}_k\};\{\boldsymbol{\psi}_k\}\right) = -N\ln\pi - N\sum_{k=1}^{K}\ln\widehat{\sigma}_k^2 - KN, \quad (12)$$

which can be minimized w.r.t. the parameters that we wish to estimate. However, the parameters of all the sources figure in (11), and estimating those at once is a difficult problem, since it is multidimensional, and highly nonlinear. A possible solution is to use an iterative procedure, such as the expectation maximization (EM) algorithm, to decompose the signal into its components, and estimate their parameters [11], [47]. For each iteration of the method, the log likelihood of the observed data $\mathbf{x}$ is increased. As shown in (4), the observed signal is modelled as a sum of M sources, where the kth channel of each inidividual source m is modelled as

$$\mathbf{x}_{k,m} = \mathbf{Z}_m\mathbf{G}_{k,m}\mathbf{a}_m + \mathbf{e}_{k,m}, \quad (13)$$

where the noise term $\mathbf{e}_k$ is decomposed into M sources, i.e.,

$$\mathbf{e}_{k,m} = \beta_m\mathbf{e}_k, \quad (14)$$

where $\beta_m \geq 0$ is chosen such that $\sum_{m=1}^{M}\beta_m = 1$. Here, β_m is chosen such that the entire error term is assigned to a single component in each iteration, i.e., $\beta_{p=m} = 1$ and $\beta_{p\neq m} = 0$, where $p = \mod(i-1, M) + 1$, with i being the EM iteration index [48], [49]. Assuming white Gaussian noise (see [11], [47]) in the E-step, the kth channel of the mth source in iteration i is modelled according to (13) based on the fundamental frequency estimate from the previous iteration, i.e.,

$$\widehat{\mathbf{x}}_{k,m}^{(i)} = \mathbf{Z}_m^{(i)}\mathbf{G}_{k,m}\widehat{\mathbf{a}}_m^{(i)} + \beta_m\left(\mathbf{x}_k - \sum_{m=1}^{M}\mathbf{Z}_m^{(i)}\mathbf{G}_{k,m}\widehat{\mathbf{a}}_m^{(i)}\right). \quad (15)$$

In the M-step, the fundamental frequency of the mth source is estimated using the nonlinear least squares (NLS) method, based on the estimate of each source from the previous iteration, i.e.,

$$\widehat{\omega}_m^{(i+1)} = \arg\min_{\omega_m}\sum_{k=1}^{K}\ln\left\|\widehat{\mathbf{x}}_{k,m}^{(i)} - \mathbf{Z}_m\mathbf{G}_{k,m}\widehat{\mathbf{a}}_m^{(i+1)}\right\|_2^2, \quad (16)$$

where the complex amplitude vector can be found, given $\widehat{\omega}_m^{(i+1)}$ as [37]

$$\widehat{\mathbf{a}}_m^{(i+1)} = \left[\sum_{k=1}^{K}\frac{\mathbf{G}_{k,m}^H\mathbf{Z}_m^H\mathbf{Z}_m\mathbf{G}_{k,m}}{\widehat{\sigma}_k^{2(i+1)}}\right]^{-1}\sum_{k=1}^{K}\frac{\mathbf{G}_{k,m}^H\mathbf{Z}_m^H\widehat{\mathbf{x}}_{k,m}^{(i)}}{\widehat{\sigma}_k^{2(i+1)}}. \quad (17)$$

The estimate of the variance σ_k^2 in iteration $i+1$ is

$$\widehat{\sigma}_k^{2(i+1)} = \frac{1}{N}\left\|\widehat{\mathbf{x}}_{k,m}^{(i)} - \mathbf{Z}_m\mathbf{G}_{k,m}\widehat{\mathbf{a}}_m^{(i+1)}\right\|_2^2. \quad (18)$$

Since the estimates of the amplitude vector and the noise variance depend on each other, we estimate the parameters in an iterative manner. If the variance of the noise is the same for both channels, the expression for the calculation of the amplitude vector becomes simpler, and it is not necessary to iterate between (17) and (18). The E- and M-steps are repeated until a convergence criterion is met, e.g., that the change in the cost function (11) is small, i.e., $J^{(i-1)} - J^{(i)} < \epsilon$ (we use $\epsilon = 10^{-6}$ in the experiments), or a maximum number of iterations have been reached (we use $I_{\text{EM}} = 10$ as upper limit on the iteration index). The method is guaranteed to converge to a local minimum, and increases the likelihood of the observed data at each step.

C. Interpretation and Fast Implementation

An approximation of the LS estimate of the amplitude vector can be obtained by noticing that the colums of $\mathbf{Z}_m$ are asymptotically orthogonal, i.e., $\lim_{N\to\infty}\mathbf{Z}_m^H\mathbf{Z}_m = N\mathbf{I}$ [11], and $\check{\mathbf{G}}_{k,m} = \mathbf{G}_{k,m}^H\mathbf{G}_{k,m} = \text{diag}(\mathbf{g}_{k,m})$, where $\mathbf{g}_{k,m} = [g_{k,m}\ \cdots\ g_{k,m}] \in \mathbb{R}^{L_m}$. The asymptotic LS amplitude estimate is

$$\check{\mathbf{a}}_m^{(i+1)} = \left[\sum_{k=1}^{K}\frac{N\check{\mathbf{G}}_{k,m}}{\check{\sigma}_k^{2(i+1)}}\right]^{-1}\sum_{k=1}^{K}\frac{\mathbf{G}_{k,m}^H\mathbf{Z}_m^H\widehat{\mathbf{x}}_{k,m}^{(i)}}{\check{\sigma}_k^{2(i+1)}}. \quad (19)$$

Furthermore, we can write the estimate of the noise variance as

$$\check{\sigma}_k^{2(i+1)} = \frac{\widehat{\mathbf{x}}_{k,m}^{(i)H}\mathbf{P}_{\mathbf{Z}_m}^{\perp}\widehat{\mathbf{x}}_{k,m}^{(i)}}{N}, \quad (20)$$

where $\mathbf{P}_{\mathbf{Z}_m}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{Z}_m}$, and,

$$\mathbf{P}_{\mathbf{Z}_m} = \mathbf{Z}_m\mathbf{G}_{k,m}\left[\sum_{k=1}^{K}\frac{N\check{\mathbf{G}}_{k,m}}{\check{\sigma}_k^{2(i+1)}}\right]^{-1}\sum_{k=1}^{K}\frac{\mathbf{G}_{k,m}^H\mathbf{Z}_m^H}{\check{\sigma}_k^{2(i+1)}} \quad (21)$$

is the orthogonal projection matrix which projects the observed signal onto the columns of $\mathbf{Z}_m\mathbf{G}_{k,m}$. The method resembles beamforming, and the panning matrix $\mathbf{G}_{k,m}$ can be interpreted as a steering matrix. We also notice that $\mathbf{Z}_m^H\widehat{\mathbf{x}}_{k,m}^{(i)}$ is the Fourier transform of $\widehat{\mathbf{x}}_{k,m}^{(i)}$ evaluated at the frequencies of the columns of $\mathbf{Z}_m^H$, and can be calculated efficiently using a fast Fourier transform (FFT). The approximate estimator can be stated as

$$\widehat{\omega}_{0,m}^{(i+1)} = \arg\min_{\omega_{0,m}}\sum_{k=1}^{K}\ln\left\|\widehat{\mathbf{x}}_{k,m}^{(i)} - \mathbf{Z}_m\mathbf{G}_{k,m}\check{\mathbf{a}}_m^{(i+1)}\right\|_2^2. \quad (22)$$

The asymptotic expression for estimating the amplitudes can be used as long as the fundamental frequencies for which it is evaluated are not very low compared to the segment length [11].

D. Initialization of the EM Algorithm

We now describe how the EM algorithm in Section III-B is initialized. Generally, EM algorithm initialization is not simple, and may result in getting stuck in a wrong local minimum. A possible approach is to use the harmonic matching pursuit (HMP) [11], [14], which is based on a residual for channel k in iteration i at time n, defined as

$$r_k^{(i)}(n) = r_k^{(i-1)}(n) - \sum_{l=1}^{L_i}g_{k,i}a_{i,l}e^{j\omega_{0,i}l(n-f_s\tau_{k,i})}. \quad (23)$$

The model parameters are estimated iteratively for each modelled harmonic source i, until a stopping criterion is met. An option is to use a detection scheme and extract sources while the residual contains harmonic components. 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 maximum a posteriori (MAP) model selection criterion [11], [50] is used as a model selection rule, i.e.,

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

where $\widehat{\mathcal{M}}_i$ is the model of the ith source, and $|\cdot|$ denotes the determinant of a matrix. The determinant of the Hessian, $\widehat{\mathbf{H}}_i$, can be approximated using the Fisher information matrix, and a normalization matrix is introduced (see [50]) 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

$$\ln\, |\widehat{\mathbf{H}}_i| = \ln\, |\mathbf{K}^{-2}| + \ln\, |\mathbf{K}\widehat{\mathbf{H}}_i\mathbf{K}|, \tag{24}$$

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 $i = 1, \ldots, I$, i.e.,

$$\left\{\widehat{\omega}_{0,i}, \widehat{L}_i\right\} = \arg\min_{\mathbf{a}_i, \{\omega_{0,i}, L_i\}} \frac{\ln\, |\mathbf{K}^{-2}|}{2} + N \sum_{k=1}^{K} \ln \left\|\boldsymbol{\beta}_{k,i}\right\|_2^2, \tag{25}$$

where $\boldsymbol{\beta}_{k,i} = \mathbf{r}_k^{(i-1)} - \mathbf{Z}_i\mathbf{G}_{k,i}\mathbf{a}_i$, and $\mathbf{r}_k^{(i)} = [r_k^i(0)\ \ r_k^i(1)\ \ \cdots\ r_k^i(N-1)]^T$. It should be noted that the cost function is multimodal, and we therefore perform the minimization with respect to $\omega_{0,i}$ using a grid search (grid size selection is discussed in [51]). The LS estimate of the amplitude vector $\mathbf{a}_i$ for each candidate $\omega_{0,i}$ are [37]

$$\widehat{\mathbf{a}}_i = \left[\sum_{k=1}^{K} \frac{\mathbf{G}_{k,i}^H\mathbf{Z}_i^H\mathbf{Z}_i\mathbf{G}_{k,i}}{\widehat{\sigma}_k^2}\right]^{-1} \sum_{k=1}^{K} \frac{\mathbf{G}_{k,i}^H\mathbf{Z}_i^H\mathbf{r}_k^{(i-1)}}{\widehat{\sigma}_k^2}, \tag{26}$$

and the estimate of noise variance in channel k is

$$\widehat{\sigma}_k^2 = \frac{1}{N}\left\|\mathbf{r}_k^{(i-1)} - \mathbf{Z}_i\mathbf{G}_{k,i}\widehat{\mathbf{a}}_i\right\|_2^2. \tag{27}$$

The fundamental frequencies and amplitudes of the M sources are then obtained by computing the residual (23) and estimating the fundamental frequency using (25) and the amplitudes using (26).

E. Harmonic Amplitude Codebooks

Taking the stereophonic mixing parameters into account when estimating fundamental frequencies simplifies the problem of estimating multiple fundamental frequencies, as described in the previous sections. However, if each submixture contains multiple sources, with distinct fundamental frequencies, estimating those is still a difficult task, especially if the fundamental frequencies and/or their harmonics are simply related, e.g., when harmonics are shared among sources. Using an iterative approach for estimating the fundamental frequencies may result in modelling several of the sources with one set of parameters, and the resulting fundamental frequency estimate is likely to be wrong (but related to the fundamental frequencies of the sources somehow). If the amplitude vector is estimated using least squares, the amplitude vector may exhibit a non-smooth amplitude envelope. To overcome this issue, we propose imposing constraints on the magnitude of the amplitude of the harmonics of each source, and to impose a form of spectral smoothness onto the sources. Specifically, we propose to train a codebook of harmonic magnitude amplitude vectors using recordings of individual instruments. The codebook is generated by jointly estimating the fundamental frequency and the model order for each frame of a set of recordings of monophonic single-source signals, and saving the corresponding complex amplitude vector. The codewords are generated by clustering vectors containing the magnitude of the entries in the complex amplitude vectors using K-means [52]. To illustrate the effectiveness of applying the principle of using a codebook of amplitude vectors when estimating multiple fundamental frequencies, consider an example in which a mixture consisting of two synthetic sources, with fundamental frequencies $f_{0,1} = 260$ Hz and $f_{0,2} = 390$ Hz, and model orders $L_1 = 10$ and $L_2 = 8$, respectively, is generated. The amplitudes of the harmonics decay as a function of the harmonic number, and the phases are randomized between 0 and 2π. The amplitude codebook in this experiment contains the true magnitude amplitude vectors. The top plot in Fig. 1 shows the magnitude spectrum of a 30 ms frame of the mixture, where the harmonics of the two signals are apparent. The bottom plot shows a comparison of cost functions where the amplitudes are estimated using least squares (NLS) and using a codebook of amplitude vectors (proposed method), respectively.

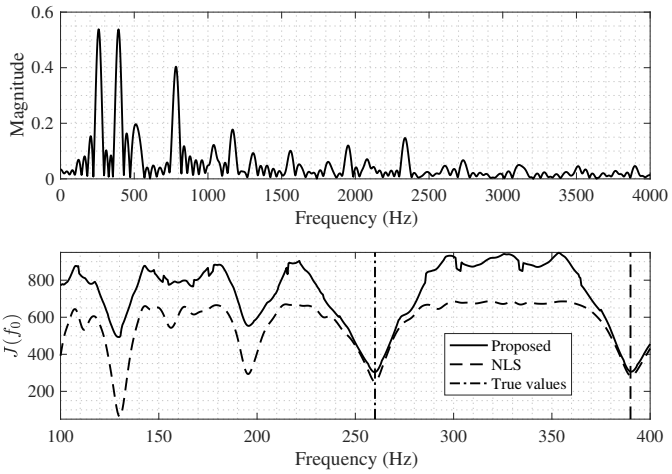

Fig. 1. Top: Magnitude spectrum of mix containig two sources with $f_{0,1} = 260$ Hz, $f_{0,2} = 390$ Hz (synthetic signals). Bottom: Value of cost function J (NLS: nonlinear least squares, Proposed: amplitude codebook used in the computation of the cost function) as a function of the candidate fundamental frequency in Hz.

In this example, the third harmonic of the first source and the second harmonic of the second source have the same frequency. This is an ill-posed problem, since we do not know how the energy is distributed across the harmonics of the sources. The amplitude vector estimated using least squares

will not fit the true signals very well. As mentioned, the magnitude values of the harmonic amplitudes for each source may be non-smooth across the harmonics, because several sources might be modelled by a more complex model, i.e., a single source where some of the harmonic components are zero. For the example above, a signal model with fundamental frequency 130 Hz would fit both sources, but the energy at the fundamental frequency, as well as the fifth, seventh and tenth harmonics, will be close to zero. In other words, the amplitude vector will be non-smooth across freqeuncy. It is also mentioned in [37] that estimating source parameters for one source at a time, results in poor amplitude estimates, if the frequncies of the components are close to each other.

The dimensions of the amplitude vectors vary with the model order and the fundamental frequency, which may vary throughout a signal. This means that we should in principle have a separate codebook for each possible number of harmonics. Since this would require a huge amount of training data, the vectors are converted to have the same dimension. A non-square transform (NST) is applied to the amplitude vectors, such that they contain the same number of entries [53]. The maximum number of harmonics, which is found as $L_{\max} = \left\lfloor \frac{2\pi}{w_{0,\min}} \right\rfloor$, where $w_{0,\min}$ is the lowest candidate fundamental frequency (in radians), is used to determine the number of rows of the NST matrix. Here, zeropadding is used, to obtain vectors of equal length, i.e., a variable-dimension amplitude vector $|\widehat{\mathbf{a}}| \in \mathbb{R}^L$ is transformed into a vector $\dot{\mathbf{a}} \in \mathbb{R}^{L_{\max}}$ of fixed dimension via the operation

$$\dot{\mathbf{a}} = \mathbf{T}_L \widehat{\mathbf{a}}, \tag{28}$$

where $\mathbf{T}_L$ is a NST matrix of dimension $L_{\max} \times L$, and is given by

$$[\mathbf{T}_L]_{i,j} = \begin{cases} 1, & \text{for } i = j. \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

The codebook $\mathcal{C}$ consists of C entries, $\{\ddot{\mathbf{a}}_c\}_{c=1}^{C}$, which are found by performing K-means clustering on the set of vectors $\{\dot{\mathbf{a}}\}$. The codebook entries are normalized to have unit norm, i.e., $\|\ddot{\mathbf{a}}_c\|^2 = 1$. When computing refined pitch and model order estimates, the cth entry in the codebook is converted to fit a dimension equal to the candidate model order L, i.e.,

$$\overline{\mathbf{a}}_c = \mathbf{T}_L^T \ddot{\mathbf{a}}_c. \tag{30}$$

The refined magnitude amplitude vector is

$$\widetilde{\mathbf{a}}_m = \underset{\gamma_m \in \mathbb{R}^+, \overline{\mathbf{a}}_c \in \mathcal{C}}{\arg\min} \|\widehat{\mathbf{a}}_m - \gamma_m \overline{\mathbf{a}}_c\|_2^2, \tag{31}$$

where γ_m is a scaling factor, to limit the size of the codebook (this is also known as gain-shape vector quantization) [31]. The resulting magnitude amplitudes in $\widetilde{\mathbf{a}}_m$ in (31) are combined with the phases of the initial amplitude estimates $\widehat{\mathbf{a}}_m$ to result in the refined estimate of the complex amplitude vector of the mth source, i.e.,

$$\mathring{\mathbf{a}}_m = [\widetilde{a}_{1,m} e^{j\angle \widehat{a}_{1,m}} \quad \cdots \quad \widetilde{a}_{L_m,m} e^{j\angle \widehat{a}_{L_m,m}}]^T. \tag{32}$$

The resulting estimate is substituted in (25), to obtain refined estimates of the fundamental frequency and model order of source m, i.e.,

$$\left\{\widehat{\omega}_{0,m}, \widehat{L}_m\right\} = \underset{\mathbf{a}_m, \{\omega_{0,m}, L_m\}}{\arg\min} \frac{\ln |\mathbf{K}^{-2}|}{2} + N \sum_{k=1}^{K} \ln \left\| \mathring{\boldsymbol{\beta}}_{k,m} \right\|_2^2, \tag{33}$$

where

$$\mathring{\boldsymbol{\beta}}_{k,m} = \mathbf{r}_k^{(m-1)} - \mathbf{Z}_m \mathbf{G}(k,m) \mathring{\mathbf{a}}_m. \tag{34}$$

Using the refined estimate of the amplitude vector in (32), the magnitude of the amplitude of each harmonic component is mapped to an entry in the codebook, which should contain smooth amplitude vectors as its entries. It should be noted that ideally we would like to have knowledge about the amplitude vectors of each source in each segment of the mixture, however, this is not possible without knowledge of the unmixed sources, and this is part of the motivation to use the amplitude codebook. However, if an amplitude vector formed using the codebook does not capture the structure of the amplitude vector corresponding to the unmixed source, errors may occur due to model mismatch, leading to erroneous fundamental frequency estimates.

F. Harmonic Component Detection

In the preliminary work presented in [38], [39], it was assumed that the number of active sources M in a frame was known a priori. However, this is often not the case, e.g., in mixtures of recordings of several musicians playing together. The musicians might play together, but all the sources might not be active at once. To overcome this issue, we propose to incorporate a detection scheme in the proposed multi-pitch estimation algorithm. It should be noted that we are here detecting harmonic sources, i.e., we consider, e.g., percussive instruments to be part of the noise component of the mixture. For simplicity, the method we derive here is for single-channel detection, however, it is straightforward to extend it, e.g., using the stereophonic signal model described in Section II. Furthermore, it should be stressed that a chord played on an instrument consists of multiple harmonic components, which we consider as separate sources. In the initial step, it is assumed that the mixture consists of $M_{\max}$ sources, with parameters estimated using the method described in Section III. We wish to determine which of the sources, parametrized by the estimates, are present in the observed signal. First, the fundamental frequency estimates $\widehat{\boldsymbol{\omega}}_0 = [\widehat{\omega}_{0,m} \dots \widehat{\omega}_{0,M_{\max}}]$ are sorted according to the likelihood, i.e.,

$$J(\widehat{\omega}_{0,m}) = \|\mathbf{x} - \mathbf{Z}_m \widehat{\mathbf{a}}_m\|_2^2, \tag{35}$$

where $\mathbf{Z}_m$ is formed using the parameter estimates $\widehat{\boldsymbol{\omega}}_0$, $\widehat{\mathbf{a}}_m = \left(\mathbf{Z}_m^H \mathbf{Z}_m\right)^{-1} \mathbf{Z}_m^H \mathbf{x}$. Furthermore, we define an ordered index $\bar{m} = \{1, \dots, \bar{M}\}$, such that $J_1 \geq \cdots \geq J_{\bar{M}}$, $\bar{\boldsymbol{\omega}} = [\omega_1 \cdots \omega_{\bar{M}}]$, and $\bar{\mathbf{L}} = [L_1 \cdots L_{\bar{M}}]$ are vectors of fundamental frequency and model order estimates, respectively, sorted according to the values of (35), and evaluated at the fundamental frequencies $\widehat{\boldsymbol{\omega}}_0$. Based on the sorted parameter estimates, we model each signal iteratively, i.e.,

$$\widehat{\mathbf{s}}_{\bar{m}} = \mathbf{Z}_{\bar{m}} \left[\mathbf{Z}_{\bar{m}}^H \mathbf{Z}_{\bar{m}}\right]^{-1} \mathbf{Z}_{\bar{m}}^H \mathbf{r}^{(\bar{m})}, \tag{36}$$

where $\mathbf{Z}_{\bar{m}}$ is formed using the corresponding fundamental frequency in $\bar{\boldsymbol{\omega}}$, and $\mathbf{r}^{(\bar{m})}$ is a residual used when estimating the complex amplitude vector of source $\bar{m}$, given by

$$\mathbf{r}^{(\bar{m})} = \mathbf{x} - \sum_{i=1}^{\bar{m}-1} \widehat{\mathbf{s}}_i. \tag{37}$$

For each of the estimated sources, we wish to determine whether it is present in the mixture or not. The method is based on the generalized likelihood ratio test (GLRT) [54], which decides between the hypoheses

$$\mathcal{H}_0 : \mathbf{r}^{(\bar{m})} = \mathbf{e} \tag{38}$$

$$\mathcal{H}_1 : \mathbf{r}^{(\bar{m})} = \mathbf{Z}_{\bar{m}}\widehat{\mathbf{a}}_{\bar{m}} + \mathbf{e}. \tag{39}$$

For each source, $\mathcal{H}_1$ (the residual contains a harmonic source) is decided if

$$L_G(\mathbf{r}^{(\bar{m})}) = \frac{p\left(\mathbf{r}^{(\bar{m})}; \widehat{\boldsymbol{\theta}}_{1,\bar{m}}, \widehat{\sigma}^2_{1,\bar{m}}\right)}{p\left(\mathbf{r}^{(\bar{m})}; \widehat{\boldsymbol{\theta}}_{0,\bar{m}}, \widehat{\sigma}^2_{0,\bar{m}}\right)} > \gamma, \tag{40}$$

where $\widehat{\boldsymbol{\theta}}_{q,\bar{m}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}_{q,\bar{m}}$ under $\mathcal{H}_q$, and γ is a threshold which determines whether $\mathcal{H}_0$ or $\mathcal{H}_1$ is chosen, and (40) can be rewritten as

$$T(\mathbf{r}^{(\bar{m})}) = \frac{N - L_{\bar{m}}}{L_{\bar{m}}}\left(L_G^{2/N}(\mathbf{r}^{(\bar{m})}) - 1\right) \tag{41}$$

$$= \frac{N - L_{\bar{m}}}{L_{\bar{m}}} \frac{\mathbf{r}^{(\bar{m})H}\mathbf{P}_{\mathbf{Z}_{\bar{m}}}\mathbf{r}^{(\bar{m})}}{\mathbf{r}^{(\bar{m})H}\mathbf{P}^{\perp}_{\mathbf{Z}_{\bar{m}}}\mathbf{r}^{(\bar{m})}} \tag{42}$$

$$= \frac{N - L_{\bar{m}}}{L_{\bar{m}}} \frac{\|\mathbf{P}_{\mathbf{Z}_{\bar{m}}}\mathbf{r}^{(\bar{m})}\|_2^2}{\|\mathbf{P}^{\perp}_{\mathbf{Z}_{\bar{m}}}\mathbf{r}^{(\bar{m})}\|_2^2} > \gamma', \tag{43}$$

where the ranks of the residual and signal subspaces for source $\bar{m}$ are $N - L_{\bar{m}}$ and $L_{\bar{m}}$, respectively, where N is the segment length, and $L_{\bar{m}}$ the model order of the source. Furthermore, $\mathbf{P}^{\perp}_{\mathbf{Z}_{\bar{m}}} = \mathbf{I} - \mathbf{P}_{\mathbf{Z}_{\bar{m}}}$ and

$$\mathbf{P}_{\mathbf{Z}_{\bar{m}}} = \mathbf{Z}_{\bar{m}}\left[\mathbf{Z}^H_{\bar{m}}\mathbf{Z}_{\bar{m}}\right]^{-1}\mathbf{Z}^H_{\bar{m}} \tag{44}$$

is the orthogonal projection matrix that projects the observed signal vector onto the columns of $\mathbf{Z}$, which is a Vandermonde matrix formed using the estimate of the fundamental frequency and the model order for the corresponding source. The resulting vector of fundamental frequency estimates is denoted $\boldsymbol{\omega}_{0,\bar{m}}$. The threshold γ' is found by choosing a desired false alarm rate P_{fa} [54], and finding the value which exceeds 1-P_{fa} samples from an F-distribution with $L_{\bar{m}}$ numerator degrees of freedom and $N - L_{\bar{m}}$ denominator degrees of freedom (see also [54, App. 9A]). A summary of the proposed method is presented in Algorithm 1.

Algorithm 1 Summary of the proposed multi-pitch estimator

Require: $\{\mathbf{x}_k\}$, $\{\mathbf{G}_{k,m}\}$, and $\mathcal{C}$.
for $m = 1, \ldots, M$ **do**
  Form initial pitch and model order estimates $\{\widetilde{\omega}^{(1)}_{0,m}, \widetilde{\mathbf{L}}^{(1)}_m\}$ by minimizing (33).
end for
for $m = 1, \ldots, M$ **do**
  while $J^{(i-1)} - J^{(i)} > \epsilon$ **and** $i < I_{\text{EM}}$ **do**
    Refine pitch estimates via EM algorithm, cf. (15) and (16), and map amplitudes to an entry in a codebook, cf. (31) and (32), resulting in $\{\widehat{\omega}^{(i)}_{0,m}\}$.
    $i \leftarrow i + 1$
  end while
end for
for $m = 1, \ldots, M$ **do** {optional}
  Perform source detection, cf. (43), resulting in $\{\widehat{\boldsymbol{\omega}}_{0,\bar{m}}\}$.
end for
return $\{\widehat{\omega}^{(i)}_{0,m}\}$ or $\{\widehat{\boldsymbol{\omega}}_{0,\bar{m}}\}$.

IV. Experiments

In this section, the performance of the proposed method is evaluated experimentally using data generated by mixing several single-instrument signals. The proposed method is evaluated in several scenarios, to validate the different parts of the method, i.e., single-channel mixtures of varying (but known) polyphony, stereophonic mixtures with varying panning parameters (known polyphony), and stereophonic mixtures with unknown polyphony. Two datasets were used to generate the mixtures used in the experiments, i.e., the IOWA database of anechoic recordings of individual instruments playing single notes[1], and the Bach10 dataset [55], which consists of multitrack recordings of ten four-part chorales composed by J. S. Bach. From the IOWA database, signals with a combined duration of 262 seconds were used to generate the mixtures. From the Bach10 database, the duration of the signals used to generate the mixtures is 1338 seconds. The dataset contains the recordings of individual instruments. In the experiments where stereophonic IOWA data is used, the panning parameters are assumed known, however, for the stereophonic Bach10 mixtures (i.e., the largest amount of data), the panning parameters of the sources are estimated using the method presented in [36], where the panning parameter distribution is modelled using a Gaussian mixture model, and the generalized variances of the Gaussian components are used to select the number of panning parameters. In all experiments, the audio signals were downsampled to $f_s = 8000$ Hz before processing. This was done to reduce the computational complexity. It should be noted that the proposed method would still work with other sample rates, and that an increase in performance might be observed with higher sample rates. The performance of the proposed method (denoted EM-CB) is compared to the performance of a parametric method for single-channel data, which does not make use of the proposed codebook-based approach (EM-LS), a state-of-the-art transcription method, which is based on a 5-dimensional dictionary of spectral templates representing the evolution of a note (BW2015) [17][2] (the standard settings were used), and the MIRtoolbox [56] implementation of the enhanced summary autocorrelation function (ESACF) [13], which is a non-parametric method for multipitch estimation. Using the MIRtoolbox, the `mirpitch()` function based on the `'Tolonen'` setting was used, however, with

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

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

`'NoFilterbank'` instead of `'2Channels'` (f_0 range similar to the proposed method).

It should be noted that for the experiments where stereophonic mixtures were used, the proposed method is implemented using the asymptotic expression for the calculation of the amplitude estimates, presented in Section III-C. However, in the evaluation of the single-channel mixtures, the exact expression in Section III is used. The reason for this is that the proposed method is computationally quite intensive, particularly in the stereophonic configuration. It should be noted that if the exact formulation in Section III-B was used, the performance would improve. With respect to computational complexity, algorithms exist for reducing the computation time (see, e.g., [51]). In all of the experiments, ground truth data is obtained for the individual harmonic sources (instrument recordings) using the joint approximate nonlinear least squares (ANLS) fundamental frequency and model order estimator from the Multi-Pitch Estimation Toolbox [11]. The performance of each of the methods is reported using metrics commonly used when evaluating multiple fundamental frequency estimation methods [57], i.e., accuracy, precision, recall and F-score, defined for each frame t as

$$a(t) = \frac{\text{TP}}{\text{FP} + \text{FN} + \text{TP}} = \frac{|\{\bar{\omega}_0\} \cap \{\widehat{\omega}_0\}|}{|\{\bar{\omega}_0\} \cup \{\widehat{\omega}_0\}|}, \quad (45)$$

$$p(t) = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{|\{\bar{\omega}_0\} \cap \{\widehat{\omega}_0\}|}{|\{\widehat{\omega}_0\}|}, \quad (46)$$

$$r(t) = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{|\{\bar{\omega}_0\} \cap \{\widehat{\omega}_0\}|}{|\{\bar{\omega}_0\}|}, \quad (47)$$

$$F(t) = \frac{2 \cdot p(t) \cdot r(t)}{p(t) + r(t)}, \quad (48)$$

respectively, along with a measure of the total error rate

$$E_{\text{total}}(t) = \frac{\max\left(|\{\bar{\omega}_0\}|, |\{\widehat{\omega}_0\}|\right) - |\{\bar{\omega}_0\} \cap \{\widehat{\omega}_0\}|}{|\{\bar{\omega}_0\}|}, \quad (49)$$

where $|\cdot|$ denotes the cardinality of a set, and $\{\bar{\omega}_0\}$ and $\{\widehat{\omega}_0\}$ are the sets of ground truth and estimated fundamental frequencies for frame t, respectively. An error is counted when an estimate deviates from the ground truth by more than half a semitone (3% deviation from the ground truth fundamental frequency). In the following subsections, we describe how the harmonic magnitude codebooks are trained, the experimental setup, present the evaluation data, and discuss the results. The parameters used in the experiments are summarized in Table I.

TABLE I
VALUES OF PARAMETERS USED IN THE EXPERIMENTS.

Parameter	Description	Value
f_s	Sampling frequency	8000 Hz
K	No. of channels	2
N	Segment length	30 ms
H	Hop size	15 ms
$f_{0,\text{min}}$	Min. pitch candidate	50 Hz
$f_{0,\text{max}}$	Min. pitch candidate	1000 Hz
L_{max}	Max. no. of harmonics	20
C	No. of codebook entries	20
M_{max}	Max. no. of sources	6
I_{EM}	Max. no. of EM iterations	10
ϵ	EM stopping criterion	10^{-6}

TABLE II
DATA FROM THE IOWA DATABASE OF INSTRUMENT RECORDINGS USED TO GENERATE THE CODEBOOKS OF HARMONIC MAGNITUDE AMPLITUDE VECTORS (V: PLAYED WITH VIBRATO).

Instrument	Instr. type	Duration (s)
Alto sax	Woodwind	61.3
Alto sax (v)	Woodwind	66.8
Bassoon	Woodwind	29.3
Bb Clarinet	Woodwind	60.8
Eb Clarinet	Woodwind	32.7
French Horn	Brass	32.9
Oboe	Woodwind	40.6
Soprano sax	Woodwind	46.9
Soprano sax (v)	Woodwind	52.3
Tenor trombone	Brass	53.5
Trumpet	Brass	103.1
Trumpet (v)	Brass	109.0
Viola	String	52.7

A. Codebook Generation

As described in Section III-E, codebooks of harmonic magnitude amplitude vectors are used to impose spectral smoothness on the amplitude vectors of the source models. The codebooks used in the experiments are trained using anechoic single-note recordings of various instruments from the IOWA database. The notes played are in the range C4-B4 (262-494 Hz), with mezzo-forte dynamics. Table II contains a description of the recordings used to train the codebooks. The recordings are processed as follows. The signals are downsampled to $f_s = 8$ kHz to decrease computation time, after which they are normalized, and processed in segments of length $N = 240$ samples (30 ms) with a hop size of $H = 120$ samples. A codebook is generated by processing individual (monophonic) anechoic recordings of instruments from the IOWA database. Each recording contains single notes played on a variety of instruments. The settings used when estimating the parameters of the signals are similar to those described in Section IV.

The fundamental frequency and the model order are estimated jointly, using a grid of fundamental frequency candidates with 1 Hz spacing, from $f_{0,\text{min}} = 100$ Hz to $f_{0,\text{max}} = f_s/2 = 4000$ Hz, in each frame of each signal. The parameters are estimated using the joint ANLS fundamental frequency and model order estimator from the Multi-Pitch Estimation Toolbox [11]. For each frame, the complex amplitude vector is estimated using a least squares approach [11], and each vector is scaled to have unit norm before performing vector quantization. Since each recording contains a succession of notes, with silence in between, only frames with power exceeding a predetermined threshold (-20 dB) are processed. As mentioned in Section III-E, the number of harmonics for a source typically varies throughout the duration of a signal. To avoid having to create a codebook for each number of harmonics (which may be impractical with a small amount of training data), the amplitude vectors are converted to a fixed dimension

using the zero-padding NST. The transformed vectors are clustered using K-means [52], and the resulting clusters are the codewords that make up the codebook. The number of entries in the codebook in these experiments has been chosen experimentally to be 20. We remark that the purpose of the experiments presented here is not to compare the performance resulting from codebooks generated differently, instead, we wish to investigate the performance of the proposed method for codebooks generated using the data in Table II.

B. IOWA Dataset Experiments

We now present the results of the evaluation of the proposed method using recordings from the IOWA database. The amplitude codebook is generated as described in Section IV-A, i.e., it consists of 20 amplitude vectors. It should be noted that the instrument recordings used to generate the test mixtures were excluded when training the codebook for the experiments presented in this section. The test signals were generated by mixing four single-channel signals each containing a single note, i.e., two recordings of a French horn and two recordings of a Bb trumpet (played with vibrato), respectively. Both instruments are played with mezzo-forte dynamics. The chords generated by mixing the individual signals are formed as follows. For root notes varying from C4 (262 Hz) to B4 (494 Hz), seven chords commonly occurring in Western music are formed, resulting in 84 chords in total, with fundamental frequencies ranging from C4 (262 Hz) to Bb5 (932 Hz). The chord types are listed using integer notation in Table III, where the notes in each chord are listed in semitone steps from the root note, i.e., a C major (7th) chord consists of the notes C (0), E (4), G (7), and B (11). The chords are generated such that none of the fundamental frequencies are integer multiples of the other fundamental frequencies, however, harmonic overlap occurs in all of the chords. In the first set of experiments, the objective is to evaluate the performance of the proposed method for single-channel mixtures containing chords generated using the IOWA data as described above, and the purpose is to evaluate the codebook approach for multi-pitch estimation. The observed mixtures have been downsampled to $f_s = 8$ kHz (to decrease the computation time), and are processed in segments of length $N = 240$ samples, with a hop size of $H = 120$ samples. Furthermore, the individual signals were normalized before creating the mixtures. The candidate fundamental frequency grid ranges from 100 Hz to 1000 Hz with 1 Hz spacing, and the maximum number of harmonics is set to $L_{\max} = 20$. The polyphony of the generated signals is four for all the chords, and this is assumed known in the experiments presented here, along with the panning parameters of the sources. The performance of the proposed method for the chord types mentioned above is shown in Table III, along with the performance of an EM algorithm, where LS amplitude estimates are used when calculating the residual in each iteration (EM-LS) [11], the transcription method presented in [17] (BW2015), and the MIRtoolbox [56] implementation of the ESACF method [13].

It can be seen from the results that the performance of the proposed method clearly outperforms the other methods on average. The mean error rates of the BW2015 and ESACF methods are approximately 43% and 59%, respectively, while for the proposed method the mean error rate is approximately 4%. The poor performance of the EM-LS method is expected, since the method suffers from issues that arise when sources share energy at their harmonics, cf. Section III-E. We can also deduce from the results that the signal model is a good fit for mixtures that contain French horn and trumpet signals. It should be noted that many of the erroneous estimates generated using the BW2015 method appear for trumpet signals which are played with vibrato. However, since vibrato occurs quite commonly, we believe that a multi-pitch estimator should be able to deal with this phenomenon.

TABLE III
ERROR RATES FOR SINGLE-CHANNEL IOWA CHORDS. FOR EACH CHORD TYPE, 12 CHORDS (ROOT NOTES C4-B4) WERE GENERATED, AND THE MEAN ERROR IS SHOWN FOR EACH CHORD TYPE (H: FRENCH HORN, T: TRUMPET PLAYED WITH VIBRATO).

Chord type	Notes (h-h-t-t)	Proposed	EM-LS	BW2015	ESACF
Maj 7	0-4-7-11	**0.0133**	0.7694	0.4945	0.4908
Min 7	0-3-7-10	**0.0335**	0.8340	0.4631	0.4873
Dom 7	0-4-7-10	**0.0111**	0.8272	0.4806	0.4729
Hdim 7	0-3-6-10	**0.0694**	0.7788	0.4281	0.5438
Dim 7	0-3-6-9	**0.0471**	0.8002	0.4712	0.6330
Min/Maj 7	0-3-7-11	**0.0244**	0.8386	0.3834	0.7012
Aug 7	0-4-8-11	**0.0765**	0.8572	0.2910	0.7686
Mean		**0.0393**	0.8150	0.4303	0.5854

Fig. 2 shows a spectrogram of a mixture of IOWA data from the experiment with amplitude and delay panning. The corresponding estimates of the fundamental frequencies in each frame of the signal are shown in the bottom part of the figure. The panning parameters were chosen such that the two submixtures, containing both horn signals and both trumpet signals, respectively, were $\theta_{\text{horn}} = 30^\circ$ and $\theta_{\text{trumpet}} = 60^\circ$, i.e., the separation angle between the submixtures is 30 degrees. The delays added to one of the channels of the submixtures were both of length 1 sample. The estimates have been sampled to make the figure easier to read. Looking at the spectrogram of the mixture, it is clear that the estimation of the fundamental frequencies is a difficult problem to solve without exploiting knowledge about the spatial information, due to harmonic overlap.

C. Bach10 Dataset Experiments

We proceed to present the experimental evaluation of the proposed method using mixtures generated using data from the Bach10 dataset, which contains the individual recordings of the parts played on a violin, clarinet, saxophone and bassoon, respectively. The amplitude codebook used in these experiments is generated as described in Section IV-A, i.e., it consists of 20 amplitude vectors, and is trained using the data in Table II. It should be noted that the mixtures generated using Bach10 data are more complex than those created using IOWA data. The recordings in the Bach10 dataset contain greater dynamic variation, since the sources are recorded with musicians playing together. Even though the instruments were

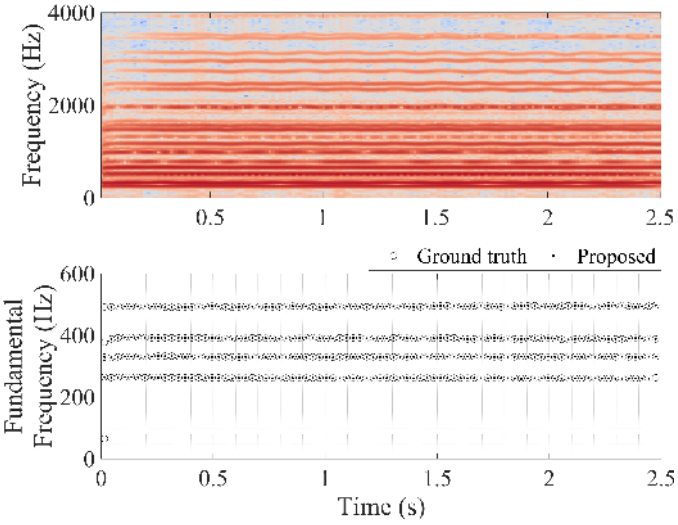

Fig. 2. Spectrogram (top) and pitch estimates (bottom) of a multi-pitch mixture of two instruments, trumpet and horn, playing the notes C4 (262 Hz), E4 (330 Hz), G4 (392 Hz) and B4 (494 Hz), respectively. Amplitude and delay panning have been applied to the two submixtures containing both horn signals and both trumpet signals, respectively.

TABLE IV
MEAN TOTAL ERROR RATES FOR SINGLE-CHANNEL MIXTURES OF VARYING POLYPHONY GENERATED FOR ALL PIECES IN THE BACH10 DATABASE (B: BASSOON, CA CLARINET, S: SAXOPHONE, V: VIOLIN).

Source(s)	Proposed (EM)	EM-LS	BW2015	ESACF
b	**0.1548**	0.0000	0.5279	0.2992
c	**0.0399**	0.0000	0.0564	0.0599
s	**0.0313**	0.0000	0.0713	0.0559
v	**0.0275**	0.0000	0.1272	0.0606
c,b	**0.2888**	0.5558	0.3078	0.4099
s,b	**0.2282**	0.3668	0.3619	0.4243
s,c	**0.1949**	0.2206	0.4294	0.4463
v,b	0.2620	0.4770	**0.0954**	0.2478
v,c	0.1554	0.3211	**0.1324**	0.3537
v,s	**0.1045**	0.1966	0.2053	0.2289
s,c,b	0.3970	0.7916	**0.2715**	0.5075
v,c,b	0.3151	0.6976	**0.3023**	0.5462
v,s,b	**0.2883**	0.6357	0.3499	0.4959
v,s,c	0.3435	0.7140	**0.1779**	0.4714
v,s,c,b	0.4099	0.7750	**0.2697**	0.6112
Mean	**0.2161**	0.3834	0.2458	0.3479

recorded separately, each of the musicians listened to the recordings of the other instruments while recording their own performance. In practice, this variation in energy results in varying signal-to-interference ratios (SIRs) for the sources.

In the first set of Bach10 experiments, the performance of the proposed method is evaluated for monophonic signals of varying, but known, polyphony, ranging from one to four. The tracks of each piece are combined uniquely with each other, resulting in 15 mixtures for each of the 10 pieces, resulting in a total of 150 different mixture of varying polyphony. The signals were downsampled to $f_s = 8$ kHz, and processed in segments of length $N = 240$ samples, with a hop size of $H = 120$ samples. The candidate fundamental frequency grid in the Bach10 experiments ranges from 50 Hz to 1000 Hz with a spacing of 1 Hz, and the maximum number of harmonics was set to $L_{\max} = 20$. For each frame of each mixture, the fundamental frequency (polyphony one) or frequencies (mixtures with polyphony greater than one) are estimated using the proposed method, an EM algorithm, where LS amplitude estimates are used when calculating the residual in each iteration (EM-LS), the transcription method presented in [17] (BW2015), and the MIRtoolbox [56] implementation of the ESACF method [13]. As with the experiments on the IOWA dataset, ground truth data was generated by jointly estimating the fundamental frequency and model order for each frame of each of the recordings of individual instruments using the joint ANLS fundamental frequency and model order estimator from the Multi-Pitch Estimation Toolbox [11]. Table IV shows the mean total error rates for the different mixtures generated by combining the tracks of the pieces in the Bach10 database, as described above.

The mean error rates for all 150 mixtures are approximately 22% for the proposed method, 25% for the BW2015 method, 35% for the ESACF method, and 35% for the EM-LS method. When looking at the results, the proposed method generally performs better than the other methods, except for some mixtures, where the error rate for the BW2015 method is lower. However, for other mixtures, the performance of the BW2015 algorithm is not very good, in particular when one of the instruments is played with vibrato. The performance of the EM-LS method is comparable to the performance of the proposed method for the single-instrument signals. This is expected since the main point of using the codebook of magnitude amplitude vectors is to increase performance for signals containing multiple sources, as described in Section III-E. For the mixtures containing multiple sources, the performance of the EM-LS method resembles the results in Section III. The performance of the proposed method for different numbers of concurrent sources indicates that it can be used for signals containing both single and multiple sources. It should be noted that in some parts of the mixtures used in these experiments, the fundamental frequencies of the sources are approximately related by integer numbers, i.e., the instruments play in unison, which makes the problem particularly difficult in these segments. In order to anticipate when poor performance is to be expected, the ratios between the ground truth fundamental frequencies for the sources should be considered. The ground truth data for a segment of one of the pieces is shown in Fig. 3.

In the next set of experiments, the performance of the proposed method is evaluated using mixtures generated by applying amplitude and delay panning to the four sources of each of the Bach10 pieces, resulting in 30 different mixtures. For each of the mixtures, four settings are used, i.e., three different sets of panning parameters, and a setting where the panning parameters are ignored when estimating the fundamental frequencies. We define $\bar{\theta} = \sum_{m=1}^{M} |\theta_m - \theta_0|$ as the sum of the absolute values of the difference between the panning direction θ_m of each source and the neutral panning position $\theta_0 = 45°$ (center). A delay of one sample was applied to the channel of each mixture component with the smallest gain applied. The polyphony of each mixture was assumed known

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878384, IEEE/ACM Transactions on Audio, Speech, and Language Processing

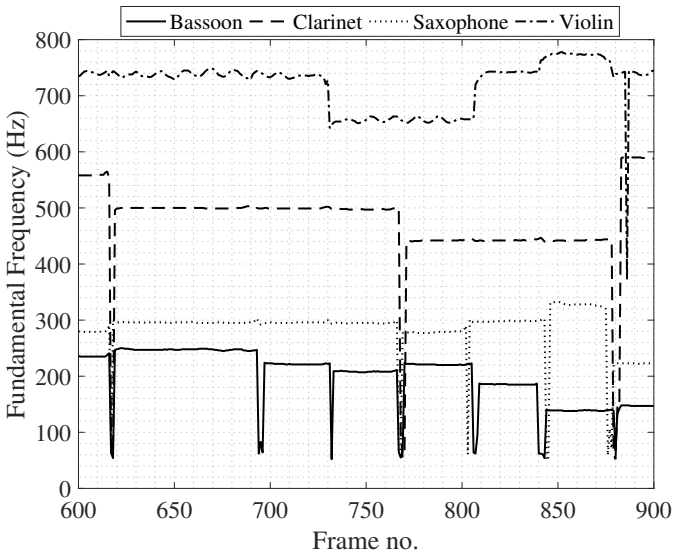

Fig. 3. Ground truth fundamental frequency estimates of a segment of the Bach chorale "Für Deinen Thron" (piece no. 8).

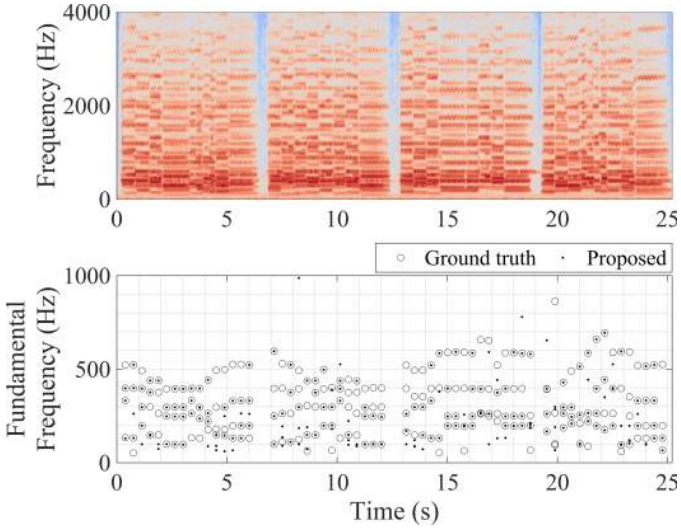

Fig. 4. Spectrogram (top) and pitch estimates (bottom) of a multi-pitch mixture of four sources from the Bach10 database with amplitude and delay panning applied to the submixtures.

TABLE V
ERROR RATES FOR STEREO BACH10 MIXTURES WITH AMPLITUDE AND DELAY PANNING APPLIED TO THE SOURCES (IN THE LAST COLUMN, THE PANNING PARAMETERS ARE IGNORED).

Piece no.	$\bar{\theta} = 50°$	$\bar{\theta} = 60°$	$\bar{\theta} = 70°$	Disregard
1	0.2789	0.2424	0.2202	0.3476
2	0.3543	0.3277	0.3113	0.4151
3	0.3193	0.2928	0.2624	0.4166
4	0.3288	0.3076	0.2822	0.4194
5	0.3805	0.3434	0.3175	0.4651
6	0.3837	0.3602	0.3298	0.4559
7	0.3056	0.2988	0.2671	0.4023
8	0.3130	0.3091	0.2781	0.3779
9	0.3722	0.3200	0.3090	0.4906
10	0.3239	0.3076	0.2826	0.3999
Mean	0.3360	0.3110	0.2860	0.4190

in these experiments, however, the panning parameters were estimated using the method presented in [36]. The results of the experiments are shown in Table V.

For the mixtures with $\bar{\theta} = 50$ ($\theta_1 = 35°$, $\theta_2 = 35°$, $\theta_3 = 60°$, $\theta_4 = 55°$), the estimated panning parameters for nine out of the ten mixtures are equal to the parameters used to generate the mixtures when rounded to integer values. For the mixture with wrongly estimated panning parameters, two panning parameters are found, and they are estimated to be between the ones used to generate the mixture. The mean error rate for the mixtures generated with these panning parameters is approximately 34%. For the mixtures with $\bar{\theta} = 60$ ($\theta_1 = \theta_2 = 30°$, $\theta_3 = \theta_4 = 60°$), the panning parameters estimated for all ten pieces are equal to the panning parameters used to generate the mixtures when rounded to the nearest integer degree and delay in samples. For one of the pieces, an extra panning parameter is found, which, however is close to the true ones. The mean error rate for the mixtures with these panning parameters is approximately 31%. For the mixtures with $\bar{\theta} = 70$ ($\theta_1 = 30°$, $\theta_2 = 25°$, $\theta_3 = 60°$, $\theta_4 = 65°$), the estimated panning parameters for eight of the mixtures are equal to the parameters used to generate the mixtures, when they are rounded to integer values. For the two mixtures with wrongly estimated panning parameters, two panning parameters are estimated, and they lie between the parameters used to generate the mixtures. For the mixtures with these panning parameters, the mean error rate is approximately 29%. To evaluate the performance of the proposed method when the panning parameters are wrongly estimated, an experiment was conducted where mixtures were generated with amplitude and delay panning, with source panning parameters $\theta_1 = \theta_2 = 30°$, $\theta_3 = \theta_4 = 60°$, respectively (the mixtures are the same as in the third column of the table), however, in the fundamental frequency estimation step, the panning parameters are ignored (the sources are wrongly assumed to be panned to the center), i.e., they are $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 45°$. The results show a consistent increase in the performance of the proposed method for all the mixtures, when the panning angles between the sources increase, and the panning parameters are estimated using the method proposed in [36]. Although the error rates increase when the panning parameters are ignored, the proposed method still works, and the error rate is approximately 42%. This result confirms that it is useful to take the panning parameters into account when estimating fundamental frequencies in stereophonic mixtures. Fig. 4 shows the spectrogram of one of the mixtures, along with the fundamental frequency estimates. The panning parameters used to generate the mixture were $\theta_1 = \theta_2 = 30°$, $\theta_1 = \theta_2 = 60°$. The data points have been sampled, to make the figure easier to read.

To evaluate the proposed detection algorithm in Section III-F, experiments have been conducted using single-channel mixtures generated using both IOWA and Bach10 data. In these experiments, the maximum number of sources is set to six, and the probability of false alarm is set to $P_{\text{fa}} = 0.05$. The IOWA mixtures considered are similar to the ones in the experiment on monophonic IOWA mixtures, i.e., consisting of

TABLE VI
PERFORMANCE OF THE PROPOSED METHOD FOR IOWA MIXTURES USING THE PROPOSED DETECTION ALGORITHM.

Chord type	Accuracy	Precision	Recall	F-measure	E_{total}
Maj 7	0.8623	0.8713	0.9885	0.9248	0.0350
Min 7	0.8583	0.8771	0.9776	0.9229	0.0438
Dom 7	0.8450	0.8563	0.9856	0.9148	0.0263
Hdim 7	0.8103	0.8629	0.9211	0.8894	0.1216
Dim 7	0.8321	0.8769	0.9350	0.9032	0.0940
Min/Maj 7	0.8398	0.8753	0.9546	0.9120	0.0765
Aug 7	0.8290	0.8775	0.9386	0.9044	0.1070
Mean	0.8395	0.8710	0.9573	0.9102	0.0720

TABLE VII
PERFORMANCE OF THE PROPOSED METHOD FOR BACH10 MIXTURES USING THE PROPOSED DETECTION ALGORITHM.

Piece	Accuracy	Precision	Recall	F-measure	E_{total}
1	0.5130	0.7022	0.6556	0.6781	0.5110
2	0.4514	0.6737	0.5777	0.6220	0.6041
3	0.4685	0.6697	0.6093	0.6380	0.5764
4	0.4090	0.6554	0.5210	0.5806	0.6354
5	0.4250	0.6759	0.5337	0.5965	0.6197
6	0.4222	0.6667	0.5352	0.5938	0.6149
7	0.5043	0.6901	0.6519	0.6705	0.6087
8	0.4596	0.7022	0.5709	0.6298	0.6075
9	0.4134	0.6344	0.5427	0.5850	0.6304
10	0.4295	0.6556	0.5546	0.6009	0.5739
Mean	0.4496	0.6726	0.5753	0.6195	0.5982

84 four-note chords. The Bach10 mixtures are generated by mixing all four tracks of each of the 10 pieces. The panning parameters for the two sets of data are the same, i.e., for one of the submixtures, the parameters are $\theta_1 = 30°$ and $\tau_1 = [0\ \ 1]^T$, while for the other submixture, the parameters are $\theta_2 = 60°$ and $\tau_2 = [1\ \ 0]^T$, respectively. For the IOWA mixtures, the panning parameters are assumed known, while the panning parameters for the Bach10 mixtures are estimated using the method presented in [36].

Table VII shows the total error rates for the mixtures. The mean error rate for the IOWA mixtures is approximately 7%, which resembles the performance of the proposed method when processing monophonic mixtures with assumed known polyphony (see Table III). For the Bach10 mixtures, the performance of the proposed method has deteriorated when compared to the performance for the monophonic mixtures with assumed known polyphony (see the bottom row in Table IV). A possible reason for the decrease in the performance of the proposed method for the Bach10 mixtures is that in many of the frames of the mixture, the fundamental frequencies of the sources are related such that they are difficult to estimate (see Fig. 3). Furthermore, as previously mentioned, the signal power of the mixture components exhibit greater variation than for the IOWA mixtures. This means that for some of the components, the signal-to-interference ratio (SIR) will be quite low, which in turn means that we can expect the performance to worsen.

V. CONCLUSION

In this paper, the problem of estimating multiple fundamental frequencies in stereophonic music mixtures is considered. Often, fundamental frequency estimation methods struggle when the components of an observed mixture are related to each other, e.g., when the harmonics of the sources coincide. To address this problem, the proposed method is based on a multi-channel harmonic signal model, in which the panning parameters of the sources in a mixture are taken into account. Furthermore, to estimate the fundamental frequencies in mixtures of sources with harmonic overlap, a codebook of amplitude vectors is trained using single-instrument recordings and used when estimating the amplitudes of the mixture components. In the experimental validation of the proposed method, the data was generated using signals from two datasets, i.e., the IOWA database of music instrument recordings, and the Bach10 database of multitrack recordings of J. S. Bach pieces. Experiments were conducted to evaluate the performance of the proposed method, which is compared to the performance of an expectation-maximization algorithm where the amplitudes of the sources are estimated using least squares (EM-LS), a multi-pitch estimator based on the enhanced summary autocorrelation function (ESACF), and a transcription method which is trained on note templates (BW2015). Different mixture configurations were used in the experiments. Monophonic mixtures of varying polyphony were used to validate the codebook approach for multiple fundamental frequency estimation for a varying number of source components. The results show that proposed method outperforms the methods to which it is compared for all the IOWA mixtures, and most of the Bach10 mixtures. To evaluate the proposed method when processing stereophonic mixtures, amplitude and delay panning was applied to source signals from the Bach10 database before mixing them. The proposed method was evaluated for both unknown panning parameters (estimated using a recently proposed method for panning parameter estimation), and erroneously estimated panning parameters. The results show an increase in performance, when the mixture components are separated via panning. When ignoring the panning parameters, the results show that the proposed method still works. The proposed method is also evaluated for monophonic mixtures with an unknown number of sources, where the proposed detection scheme (see Section III-F) is used to estimate the polyphony. For the IOWA mixtures used in this part of the evaluation, the results are similar to those for monophonic mixtures with known polyphony, while for the Bach10 mixtures the performance has decreased. A possible reason for the decrease in the performance is that the components of the Bach10 mixtures exhibit greater variation in signal power than the components of the IOWA mixtures. Overall, the results indicate that the proposed method outperforms the transcription methods and multipitch estimators to which they are compared. The extension of the harmonic signal model to multiple channels, and the restriction on the amplitudes of the harmonic components of each source via a codebook of amplitude vectors is relatively straightforward, which is one of the principal advantages of

this type of method. For non-parametric methods, such modifications may be more difficult. Future work includes extending the signal model to allow, e.g., inharmonicity, investigating different ways of imposing smoothness on the amplitudes of the harmonic components (e.g., different ways of training the amplitude codebooks), and testing the proposed methods in some of the applications mentioned in Section I.

REFERENCES

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

[2] K. Kokkinakis and P. C. Loizou, *Advances in Modern Blind Signal Separation Algorithms: Theory and Applications*, ser. Synthesis Lectures on Algorithms and Software in Engineering. Morgan & Claypool Publishers, 2010.

[3] G. R. Naik and W. Wang, *Blind Source Separation: Advances in Theory, Algorithms and Applications*. Springer Publishing Company, Incorporated, 2014.

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

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul 2002.

[6] D. Giannoulis and A. Klapuri, "Musical instrument recognition in polyphonic audio using missing feature approach," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1805–1817, Sept 2013.

[7] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sept 2012.

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

[9] 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.

[10] D. Talkin, *A robust algorithm for pitch tracking (RAPT)*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B.V., 1995.

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

[12] M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr 1990, pp. 357–360 vol.1.

[13] 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.

[14] 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.

[15] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, 2003, pp. 177–180.

[16] 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.

[17] 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.

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

[19] 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.

[20] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 528–537, March 2010.

[21] F. Elvander, T. Kronvall, S. Adalbjörnsson, and A. Jakobsson, "An adaptive penalty multi-pitch estimator with self-regularization," *Signal Process.*, vol. 127, pp. 56 – 70, 2016.

[22] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[23] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing. Springer, 2008.

[24] F. Flego and M. Omologo, "Robust f0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *Proc. European Signal Processing Conf.*, 2006.

[25] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 409–412, 2012.

[26] V. Pulkki, *Spatial sound generation and perception by amplitude panning techniques*. Helsinki University of Technology, 2001.

[27] B. Katz, *Mastering Audio - The Art and the Science*, E. James, Ed. Focal Press, 2007.

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

[29] H. Haas, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc.*, vol. 20, no. 2, pp. 146–159, 1972.

[30] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing - Processing and Perception of Speech and Music, Second Edition.* Wiley, 2011.

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

[32] 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.*, vol. 5, May 2006, pp. V–V.

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

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

[35] 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.

[36] J. M. Hjerrild and M. G. Christensen, "Estimation of source panning parameters and segmentation of stereophonic mixtures," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018.

[37] 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.

[38] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Estimation of multiple pitches in stereophonic mixtures using a codebook-based approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017.

[39] ——, "Multi-pitch estimation of audio recordings using a codebook-based approach," in *Proc. European Signal Processing Conf.*, 2016.

[40] ——, "Pitch estimation of stereophonic mixtures of delay and amplitude panned signals," in *Proc. European Signal Processing Conf.*, 2015.

[41] M. W. Hansen, J. M. Hjerrild, M. G. Christensen, and J. Kjeldskov, "Parametric multi-channel separation and re-panning of harmonic sources," *Proc. Int. Conf. Digital Audio Effects*, 2018.

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

[43] F. Elvander, S. I. Adalbjörnsson, J. Karlsson, and A. Jakobsson, "Using optimal transport for estimating inharmonic pitch signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, March 2017, pp. 331–335.

[44] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, ser. Signals and Communication Technology. Springer, 2010.

[45] S. Gorlow and J. D. Reiss, "Model-based inversion of dynamic range compression," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1434–1444, 2013.

[46] K. Kim and G. Shevlyakov, "Why gaussianity?" *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 102–113, March 2008.

[47] 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.

[48] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the em algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, April 1993, pp. 728–731 vol.2.

[49] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2664–2677, Oct 1994.

[50] 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.
[51] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188 – 197, 2017.
[52] 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.
[53] 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.
[54] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, Inc., 1998.
[55] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 2121–2133, Nov 2010.
[56] 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.
[57] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, October 26-30 2009, pp. 315–320.