PARAMETRIC MULTI-CHANNEL SEPARATION AND RE-PANNING OF HARMONIC SOURCES

M. W. Hansen †, *J. M. Hjerrild* †, *M. G. Christensen* †

†Audio Analysis Lab, CREATE
Aalborg University
Aalborg, Denmark
{mwh,jmhh,mgc}@create.aau.dk

J. Kjeldskov *

*Department of Computer Science
Aalborg University
Aalborg, Denmark
jesper@cs.aau.dk

ABSTRACT

In this paper, a method for separating stereophonic mixtures into their harmonic constituents is proposed. The method is based on a harmonic signal model. An observed mixture is decomposed by first estimating the panning parameters of the sources, and then estimating the fundamental frequencies and the amplitudes of the harmonic components. The number of sources and their panning parameters are estimated using an approach based on clustering of narrowband interaural level and time differences. The panning parameter distribution is modelled as a Gaussian mixture and the generalized variance is used for selecting the number of sources. The fundamental frequencies of the sources are estimated using an iterative approach. To enforce spectral smoothness when estimating the fundamental frequencies, a codebook of magnitude amplitudes is used to limit the amount of energy assigned to each harmonic. The source models are used to form Wiener filters which are used to reconstruct the sources. The proposed method can be used for source re-panning (demonstration given), remixing, and multi-channel upmixing, e.g. for hi-fi systems with multiple loudspeakers.

1. INTRODUCTION

Music signals often contain a mixture of multiple instrument recordings. To process such a mixture, e.g., with the goal of modifying the sources independently, it may be beneficial to extract the individual sources in the mixture. This task is known as source separation, and it has applications in areas such as music information retrieval [1], sound scene modification [2], and enhancement [3].

The problem of separating sources in a music mixture is in general very difficult, because of the presence of overlap in both time and frequency. In such cases, the source separation problem is in many cases ill-posed, and the single-channel source separation problem is very difficult to solve, and would rely heavily on prior information about the sources. When multiple channels of data are available, it is possible to exploit information about the mixing process. A method for separating two sources from a single-channel mixture was proposed in [4], based on a sparse non-negative decomposition algorithm, whereas in [5] a method based on single-channel non-negative matrix factorization (NMF) was proposed for polyphony music transcription. In [6], a method based on non-negative matrix factorization (NMF) for stereophonic source separation is presented, while in [7] a framework for incorporating prior knowledge in source separation is presented. Separation of moving sources is considered in [8] using a method based on multi-channel NMF. Time-variation is allowed through the use of spatial covariance matrices (SCMs) which are generated based on estimated directions of arrival (DOAs). Separation of sources from multi-channel reverberant mixtures, although in a semi-blind fashion, with known mixing filters, was considered in [9]. Re-panning of stereophonic sources was proposed in [10] for a known number of sources without delay panning.

Parametric signal models, where the sinusoidal components of a signal are modelled as a sum of sinusoids, can also be used for source separation. A method for source separation and auditory scene analysis based on a multi-pitch and periodicity analysis method is presented in [11], while sinusoidal modelling was used for separating harmonic sources using a classification method to group extracted sinusoids in [12]. Spectral overlap often occur in music signals, and this should be taken into account when estimating the parameters of the sources. A source separation method based on pitch, amplitude modulation and spatial cues for separation of harmonic instruments from stereo music recordings is proposed in [13]. In [14], a method for reconstruction of completely overlapped notes is presented, where the spectral envelope of each source is learnt in segments without overlap, and then used to extract the sources. A separation approach based on optimal filtering is presented in [15], where a linearly constrainted minimum variance (LCMV) filter is constructed based on a priori knowledge in the form of score information. Furthermore, the balance between overlapping harmonics is adjusted using a priori knowledge about the magnitude of each harmonic.

In this paper, we present a method for extracting harmonic sources from stereophonic mixtures of music recordings, such as those made artificially in a studio. First, the panning parameters and activations of the sources are estimated using a method based on clustering of narrowband interaural level and time differences (ILDs and ITDs) (see [16] for further details). Usually, in source separation algorithms, the number of sources is assumed known a priori (see, e.g., [6]), however, here the number of sources does not need to be known. Equipped with the estimated panning parameters, the fundamental frequencies of the harmonic sources are estimated, along with the number of harmonics, and the harmonic amplitudes, using an iterative approach. To enforce spectral smoothness, a codebook of magnitude amplitudes trained on recordings of harmonic sources is used (see [17] for further details). The source models are used to form a Wiener filter for extraction of each source from the mixture. It should be noted that the proposed method is also capable of separating sources from monophonic, i.e., single-channel mixtures. After the sources have been extracted, they are combined with new panning parameters, and the residual, i.e., the parts of the mixture not captured by the harmonic model of the sources.

* Supported by the Technical Faculty of IT and Design, Aalborg University.

2. SIGNAL MODEL

An observed multichannel mixture is modelled as a sum of M harmonic sources s_m, $m = 1, \ldots, M$, plus a noise term e. The signal in the kth channel at time n is

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

where $g_{k,m}$ and $\tau_{k,m}$ are the amplitude and delay panning parameters, respectively. An example of an amplitude panning law, which could used to calculate the gains applied to each channel of a stereophonic mixture is [18]

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

where $\phi_m \in [0, \text{pi}/2]$. The mth source s_m is modelled as a sum of L_m harmonic components, i.e,

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

where $\omega_{0,m}$ is the fundamental frequency of the mth source, L_m is the model order, and $\alpha_{m,l} = A_{m,l} e^{j\phi_{m,l}}$ is the complex amplitude of the lth harmonic, where $A_{m,l}$ is the real amplitude and $\phi_{m,l}$ its phase. A complex signal model is used because it may result in simplified expressions, and a lower computational complexity. The signal model may be used with real signals by applying the Hilbert transform. It should be noted that although we focus on the stereophonic case ($k = 2$), we here present a general multi-channel signal model, which can be used in scenarios where $k > 2$ using a different panning law. Furthermore, according to the source model (3), an instrument recording may contain multiple sources, e.g., when a chord is played on a guitar, where the signal generated by each string is considered to be a source. Furthermore, we define a submixture as a sum of sources that share panning parameters. The kth channel of an observed mixture is processed in segments each containing N consecutive samples, i.e,

$$\mathbf{x}_k = [x_k(0) \;\; x_k(1) \;\; \cdots \;\; x_k(N-1)]^T, \tag{4}$$

which can be used to write the signal model in vector form as

$$\mathbf{x}_k = \sum_{m=1}^{M} \mathbf{Z}_m \mathbf{G}_{k,m} \boldsymbol{\alpha}_m + \mathbf{e}_k, \tag{5}$$

where $\mathbf{Z}_m$ is a Vandermonde matrix, with the harmonic components of source m 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 (2) and $\tau_{k,m}$ for channel k of source m, i.e.,

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

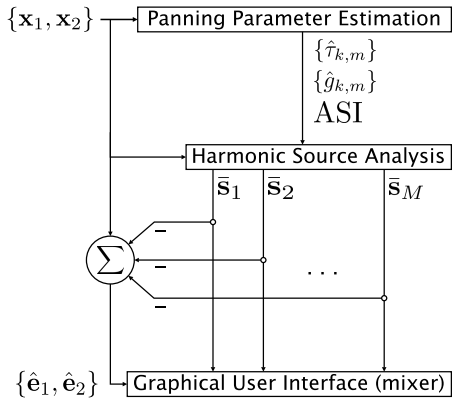

Figure 1: Overview of the proposed method.

When only amplitude panning is applied, $\tau_{k,m} = 0 \;\forall\; \{k, m\}$, 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 signal. The vector of complex amplitudes for source m is given by

$$\boldsymbol{\alpha}_m = [\alpha_{m,1} \;\; \cdots \;\; \alpha_{m,L_m}]^T, \tag{6}$$

and the noise vector is

$$\mathbf{e}_k = [e_k(0) \;\; e_k(1) \;\; \cdots \;\; e_k(N-1)]^T. \tag{7}$$

Since we model the sinusoidal source components, the noise term contains the non-periodicities that are not captured by the harmonic model. In the next section, we present the proposed method for estimating the panning parameters $g_{k,m}$ and $\tau_{k,m}$, along with the number of unique panning parameters, which corresponds to the number of submixtures.

3. PROPOSED METHOD

The proposed method consist of several sub-systems, as shown in Figure 1. In the initial step of the proposed method, the panning parameters of the sources in the mixture are estimated, along with an active source indication (ASI) of when the corresponding sources are active. This knowledge is exploited in the harmonic source analysis, where the parameters of each source $\mathbf{s}_m$ in the mixture are estimated, i.e., its fundamental frequency $\omega_{0,m}$, the number of harmonics L_m, and the amplitude vector $\boldsymbol{\alpha}_m$. The harmonic models of the sources are used to form Wiener filters, which are used to extract the sources from the mixture. The resulting frames are combined using overlap-add, and a graphical user interface (GUI) is used to re-pan the sources.

3.1. Panning Parameter Estimation and Activity Detection

As shown in Figure 1, the panning parameters of the sources in the observed multi-channel mixture are required as input for the proposed harmonic signal analysis sub-system. The source panning parameters are estimated along with the number of unique panning parameters using the method presented in [16]. The method is a blind source panning estimation algorithm based on clustering of narrowband interaural level and time differences (ILDs, ITDs). For an unknown number of sources, the parameter distribution

across all segments of the mixture is modelled as a Gaussian mixture. The generalized variance and degree of membership of the Gaussian components across segments are used as a basis for the selection of clusters amongst candidates. In the time-frequency domain we define a vector $\mathbf{y}$ for each frame containing the relative amplitude panning parameters and relative channel delays, i.e.,

$$\mathbf{y}=\Big[\hat{g}(\omega),\hat{\tau}(\omega)\Big]^T=\bigg[\arctan\bigg(\bigg|\frac{X_1(\omega)}{X_2(\omega)}\bigg|\bigg),\frac{1}{\omega}\angle\frac{X_2(\omega)}{X_1(\omega)}\bigg]^T, \quad (8)$$

where $\hat{\tau}(\omega)=\hat{\tau}_1(\omega)-\hat{\tau}_2(\omega)$, $X_k(\omega)$ is the discrete Fourier transform of the kth channel of a segment of the mixture, and $\angle$ denotes phase. Eq. (8) is constrained on the assumption of W-disjoint orthogonality [19] and on the so-called narrowband assumption that requires the maximum frequency ω_{max} and maximum delay τ_{max} to be strictly within the range $|\omega_{\text{max}}\tau_{\text{max}}|<\pi$. From (8) we collect P observations $\mathcal{Y}=\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(P)}\}$ with identical probability distributions, each being mutually independent. The log-likelihood function of the P observations is

$$\ln p(\mathcal{Y}|\boldsymbol{\theta})=\sum_{p=1}^{P}\ln\sum_{m=1}^{M}\gamma_m p(\mathbf{y}^{(p)}|\boldsymbol{\theta}_m), \quad (9)$$

where $\boldsymbol{\theta}_m$ is the unknown and deterministic parameter vector of the mth source. For the purpose of estimating panning parameters, the distribution of $\mathbf{y}$ from Eq. (8) is modelled as a Gaussian mixture of M sources, with diagonal covariance matrices, i.e.,

$$p(\mathbf{y}|\boldsymbol{\theta})=\sum_{m=1}^{M}\gamma_m\frac{\exp\big\{-\frac{1}{2}(\mathbf{y}-\mu_m)^T\mathbf{C}_m^{-1}(\mathbf{y}-\mu_m)\big\}}{\sqrt{(2\pi)^d\det(\mathbf{C}_m)}}, \quad (10)$$

where $\boldsymbol{\theta}\triangleq\{\gamma_1,\ldots,\gamma_M,\hat{\boldsymbol{\mu}}_1,\ldots,\hat{\boldsymbol{\mu}}_M,\mathbf{C}_1,\ldots,\mathbf{C}_M\}$ is the complete set of parameters, where the set $\{\gamma_m,\hat{\boldsymbol{\mu}}_m,\mathbf{C}_m\}$ denotes the mixing probability, mean and covariance of the mth Gaussian. In general, $\gamma_m\geq 0,\quad\sum_{m=1}^{M}\gamma_m=1$, for $m=1,\ldots,M$. The maximum likelihood (ML) estimate of the parameter vector is

$$\hat{\boldsymbol{\theta}}_{\text{ML}}=\arg\max_{\boldsymbol{\theta}}\ \ln p(\mathcal{Y}|\boldsymbol{\theta}) \quad (11)$$

for a value of M such that the GMM is overfitted, see [16]. The ML GMM parameter estimates in $\hat{\boldsymbol{\theta}}_{\text{ML}}$ are obtained using an EM-algorithm. Several GMM EM-methods have been proposed for estimating the number of sources, using a penalty term such as the Bayesian information criterion (BIC) or the minimum description length (MDL) [20]. However, the problem is complicated for audio recordings for two reasons: no unique definition of a "true cluster" necessarily exists, and the assumption of normality does not exactly hold, see, e.g., [21]. Therefore, each of the underlying GMM components does not necessarily correspond to a source cluster.

In the present method clusters are selected among Gaussian component candidates by fitting a GMM to the observed data with a large number of components. From the overfitted GMM clusters are defined as having lowest generalized variance δ and as being well separated from other candidates as described in the following. The cluster indices are columns of $\zeta_{\omega s}$ which have low generalized variance δ, and are well separated from all GMM components, and $\hat{\boldsymbol{\theta}}_s$ is arranged such that $\delta_1<\delta_2<\cdots<\delta_S$, where $\delta=\det(\hat{\mathbf{C}})$ and $s=\{1,2,\cdots,S\}$. The a posteriori probability $\zeta_{\omega s}$ that $\mathbf{y}_\omega$ belongs to mixture component s is

$$\zeta_{\omega s}=\frac{\hat{\gamma}_s\mathcal{N}(\mathbf{y}_\omega|\hat{\boldsymbol{\mu}}_s,\hat{\mathbf{C}}_s)}{\sum_{j=1}^{S}\hat{\gamma}_j\mathcal{N}(\mathbf{y}_\omega|\hat{\boldsymbol{\mu}}_j,\hat{\mathbf{C}}_j)}. \quad (12)$$

The sth column does not represent a cluster if $0<\zeta_{\omega\epsilon}<1\wedge 0<\zeta_{\omega s}<1$, where $\epsilon=\{1,2,\cdots,s-1\}\,\forall\,\omega$. After ranking, the $\hat{M}$ clusters are in the first columns of $\zeta_{\omega s}$, as observed in [16]. This leads to an estimate of the M unique panning parameters and the statistics $\hat{\boldsymbol{\theta}}_{\hat{M}}$ from which the vector $\hat{\boldsymbol{\mu}}_m$ is the panning parameters of the mth source, across all segments.

We compute an active source indication (ASI) for each frame of the observed mixture. Specifically, the input signal is processed in frames of length 60 ms, with a hop size of 15 ms. In each frame all possible combinations of the obtained $\hat{\boldsymbol{\theta}}_{\hat{M}}$ statistics are fitted to the observed data $\mathbf{y}$ resulting in a new GMM likelihood. The maximum likelihood combination is chosen for each frame. The obtained ASI is a binary indication of activity of each panning parameter in each frame of the mixture, and is used as input to the harmonic analysis sub system.

3.2. Harmonic Signal Analysis

In this section the method used to analyse the harmonic sources in a stereophonic mixture is presented. The goal is to estimate the fundamental frequencies of the harmonic components in the mixture, along with the number of harmonics for each source, and the complex amplitudes, provided with information about the source panning parameters, and source activity indication, as described in the previous section. The proposed method is based on the maximum likelihood principle, and the log-likelihood of the kth channel of an observed signal is parametrized by $\boldsymbol{\psi}_k=[\boldsymbol{\psi}_{k,1}\ \cdots\ \boldsymbol{\psi}_{k,M}]^T$, where $\boldsymbol{\psi}_{k,m}=[\omega_{0,m}\ g_{k,m}\ \tau_{k,m}\ \boldsymbol{\alpha}_m^T]^T$, 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 different variance in each channel, σ_k^2. Defining the error as $\mathbf{e}_k=\mathbf{x}_k-\sum_{m=1}^{M}\mathbf{Z}_m\mathbf{G}_{k,m}\boldsymbol{\alpha}_m$, the likelihood of the kth channel of the observed signal is defined as

$$p\left(\mathbf{x}_k;\boldsymbol{\psi}_k\right)=\frac{1}{\left(\pi\sigma_k^2\right)^N}e^{-\frac{1}{\sigma_k^2}\|\mathbf{e}_k\|_2^2}, \quad (13)$$

which across channels becomes

$$p\left(\{\mathbf{x}_k\};\{\boldsymbol{\psi}_k\}\right)=\prod_{k=1}^{K}\frac{1}{\left(\pi\sigma_k^2\right)^N}e^{-\frac{1}{\sigma_k^2}\|\mathbf{e}_k\|_2^2}. \quad (14)$$

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

$$\ln p\left(\mathbf{x}_k;\boldsymbol{\psi}_k\right)=-N\ln\pi-N\ln\sigma_k^2-\frac{\|\mathbf{e}_k\|_2^2}{\sigma_k^2} \quad (15)$$

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

$$\ln p\left(\{\mathbf{x}_k\};\{\boldsymbol{\psi}_k\}\right)=-KN\ln\pi-N\sum_{k=1}^{K}\ln\sigma_k^2-\sum_{k=1}^{K}\frac{\|\mathbf{e}_k\|_2^2}{\sigma_k^2}. \quad (16)$$

The fundamental frequencies, complex amplitudes, and noise variance for each channel are estimated by maximizing (16). Since the problem of estimating the parameters of all the sources at once is impractical in terms of computational complexity, the parameters are estimated iteratively using an EM algorithm. For each iteration of the method, the log-likelihood of the observed segment of the mixture is increased. The observed signal is modelled as a sum of M sources, where the kth channel of source m is modelled as

$$\mathbf{x}_{k,m}=\mathbf{Z}_m\mathbf{G}_{k,m}\boldsymbol{\alpha}_m+\mathbf{e}_{k,m}, \quad (17)$$

where $\mathbf{G}_{k,m}$ is now formed using the estimates $\{\hat{g}_{k,m}, \hat{\tau}_{k,m}\}$ for each source, and where the noise term $\mathbf{e}_k$ is decomposed into M sources, i.e.,

$$\mathbf{e}_{k,m} = \beta_m \mathbf{e}_k, \tag{18}$$

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 $\beta_{p\neq m} = 0$, and $p = \mod(i-1, M) + 1$, with i being the EM iteration index [22, 23]. Assuming white Gaussian noise (see [24, 25]), in the E-step, the kth channel of the mth source in iteration i is modelled according to (17) based on parameters estimated in the previous iteration, i.e.,

$$\hat{\mathbf{x}}_{k,m}^{(i)} = \mathbf{Z}_m^{(i)} \mathbf{G}_{k,m} \hat{\boldsymbol{\alpha}}_m^{(i)} + \beta_m \left(\mathbf{x}_k - \sum_{m=1}^{M} \mathbf{Z}_m^{(i)} \mathbf{G}_{k,m} \tilde{\boldsymbol{\alpha}}_m^{(i)} \right), \tag{19}$$

where $\widetilde{\boldsymbol{\alpha}}_m = [\widetilde{A}_{1,m} e^{j\angle\widehat{\alpha}_{1,m}} \;\; \cdots \;\; \widetilde{A}_{L_m,m} e^{j\angle\widehat{\alpha}_{L_m,m}}]^T$ is formed using a scaled codebook entry $\tilde{\mathbf{A}}_m$ from a codebook $\mathcal{C}$ of magnitude amplitude vectors trained on individual notes played on a variety of instruments, and combined with the phases resulting from the least squares estimate of the complex amplitude vector, given $\hat{\omega}_m^{(i+1)}$ as [26] (see [17] for more information)

$$\hat{\boldsymbol{\alpha}}_m^{(i+1)} = \left[\sum_{k=1}^{K} \frac{\mathbf{G}_{k,m}^H \mathbf{Z}_m^H \mathbf{Z}_m \mathbf{G}_{k,m}}{\hat{\sigma}_k^{2(i+1)}} \right]^{-1} \sum_{k=1}^{K} \frac{\mathbf{G}_{k,m}^H \mathbf{Z}_m^H \hat{\mathbf{x}}_{k,m}^{(i)}}{\hat{\sigma}_k^{2(i+1)}}. \tag{20}$$

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

$$\hat{\omega}_m^{(i+1)} = \arg\min_{\omega_m} \sum_{k=1}^{K} \ln \left\| \hat{\mathbf{x}}_{k,m}^{(i)} - \mathbf{Z}_m \mathbf{G}_{k,m} \tilde{\boldsymbol{\alpha}}_m^{(i+1)} \right\|_2^2, \tag{21}$$

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

$$\hat{\sigma}_k^{2(i+1)} = \frac{1}{N} \left\| \hat{\mathbf{x}}_{k,m}^{(i)} - \mathbf{Z}_m \mathbf{G}_{k,m} \tilde{\boldsymbol{\alpha}}_m^{(i+1)} \right\|_2^2. \tag{22}$$

The complex amplitude vector and the noise variance are estimated in an iterative fashion, because they depend on each other. It is not necessary to iterate between (20) and (22) if the noise variance for both channels are equal. The E- and M-steps are repeated until a convergence criterion is met. The method is guaranteed to converge to a local minimum in each step, and increases the likelihood of the observed data at each step. Initialization of the EM algorithm is not simple, and can result in poor performance, if it is not done carefully. We here use the harmonic matching pursuit (HMP) [27, 24], which is based on a residual for channel k in iteration m at time n, defined as

$$r_k^{(m)}(n) = r_k^{(m-1)}(n) - \sum_{l=1}^{L_m} g_{k,m} \alpha_{m,l} e^{j\omega_{0,m} l (n - \tau_{k,m})}. \tag{23}$$

The model parameters are estimated iteratively for each modelled 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 maximum a posteriori (MAP) model selection criterion [28, 24] is used as a model selection rule, i.e.,

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

where $\hat{\mathcal{M}}_m$ is the model of the mth source, and $|\cdot|$ denotes the determinant of a matrix. The determinant of the Hessian, $\hat{\mathbf{H}}_m$, can be approximated using the Fisher information matrix, and a normalization matrix is introduced (see [28]) such that

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

where the last term, which is of order $\mathcal{O}(1)$, is ignored, and the first term is used as a penalty term (see [17] for more details). We can now state the joint pitch and model order estimator used to compute initial estimates for sources $m = 1, \ldots, M$, i.e.,

$$\left\{\hat{\omega}_{0,m}, \hat{L}_m\right\} = \arg\min_{\boldsymbol{\alpha}_m, \{\omega_{0,m}, L_m\}} \frac{\ln |\mathbf{K}^{-2}|}{2} + N \sum_{k=1}^{K} \ln \left\|\boldsymbol{\beta}_{k,m}\right\|_2^2, \tag{25}$$

where

$$\boldsymbol{\beta}_{k,m} = \mathbf{r}_k^{(m-1)} - \mathbf{Z}_m \mathbf{G}_{k,m} \tilde{\boldsymbol{\alpha}}_m, \tag{26}$$

and $\mathbf{r}_k^{(m)} = [r_k^m(0) \;\; r_k^m(1) \;\; \cdots \;\; r_k^m(N-1)]^T$. Since the cost function is multi-modal, it is minimized with respect to $\omega_{0,m}$ using a grid search (grid size selection is discussed in [29]). The fundamental frequencies and amplitudes of the M sources are obtained by iterating between the expectation and maximization steps, i.e., (19), and (20)-(22), respectively, until convergence.

3.3. Source Reconstruction and Re-Panning

The harmonic sources in an observed stereophonic mixture are implicitly modelled in the iterative parameter estimation process, i.e., the estimate of the mth source is

$$\hat{\mathbf{s}}_m(n) = \mathbf{Z}_m(n) \hat{\boldsymbol{\alpha}}_m, \tag{27}$$

for $n = 1, \ldots, N$. Since the number of entries in the amplitude codebook $\mathcal{C}$ is relatively small, the signals $\hat{\mathbf{s}}_m$, for $m = 1, \ldots, M$, may sound a bit rough when listened to directly. Instead, we propose to use the estimated parameters to form a frequency-domain Wiener filter to extract each source from a segment of the observed mixture, i.e.,

$$\bar{S}_m(\omega) = \frac{\|\hat{S}_m(\omega)\|^2}{\|\hat{S}_m(\omega)\|^2 + \|\hat{V}(\omega)\|} X(\omega), \tag{28}$$

where $\bar{S}_m(\omega)$ is the frequency-domain filter output at a certain frequency bin corresponding to ω, $\hat{S}_m(\omega)$ is the DFT of the source estimate $\hat{\mathbf{s}}_m$, $\hat{V}(\omega)$ is the DFT of the estimates of the interfering sources and the noise, i.e., $\mathbf{v} = \mathbf{x} - \hat{\mathbf{s}}_m$, $X(\omega)$ is the DFT of a single-channel version of the mixture. Each time-domain segment of each the M sources is generated as the inverse DFT of the filtered output above. The segments are combined using overlap-add.

4. EXPERIMENTS

The experimental evaluation of the proposed method for panning parameter estimation, source separation and re-panning consists of multiple experiments. To evaluate the performance of the proposed method for source separation, a multitrack recording from the MedleyDB database of music recordings [30] is used, i.e., Aimee Norwich - Flying. A segment containing 24 seconds (start: 105.5 s, end: 129.5 s) of audio from three instrument recordings

Table 1: Description of the data used in the experiments.

File name (.wav)	Instrument	ϕ (degrees)	τ (samples)
Flying_RAW_14_01	Trombone	30	0
Flying_RAW_03_02	Bass	5	0
Flying_RAW_15_02	Clarinet	-30	0

Table 2: Pannning parameter estimates.

Track	$\hat{\phi}$ (degrees)	$\hat{\tau}$ (samples)
Trombone	29.99	0.00
Bass	4.99	0.01
Clarinet	-29.97	0.00

are amplitude panned to synthetically generate a stereophonic mixture. Descriptions of the tracks used in the mixture and their panning parameters are presented in Table 1.

The estimation of the number submixtures and their panning parameters are evaluated on the observed stereo mixture with $f_s = 44.1$ kHz. The input signal is processed in samples of length $N = 2640$ samples (60 ms), with a hop size of $H = 662$ samples (15 ms). The GMM is overfitted with $M = 10$ and from the overfitted GMM components, an estimate of the source clusters are obtained. To lower the computational complexity and remove part of the noise floor from the spectrum, we select the frequency bins in the measurement vector (8) according to an indicator function $b(\omega)$ defined for all ω, i.e.,

$$b(\omega) = \begin{cases} 1, & |X_1(\omega)||X_2(\omega)| > |\mathbf{X}_1|^T|\mathbf{X}_2|/N \\ 0, & \text{otherwise} \end{cases} . \quad (29)$$

The estimated source clusters are shown in Figure 3. The source panning clusters are visualized, as overlayed on the data and $\mathbf{y}$, and the contours of the initial overfitted GMM components. Both amplitude panning angle and delay were estimated correctly and the results are shown in table 2. We observe that the panning parameters are almost equal to the true parameters. The number of sources has been estimated to the true value of $M = 3$. Next, we can evaluate the ASI estimation shown in Figure 2. The Figure shows the ASI overlayed on the unmixed sources. A black vertical line indicates activity in the given frame at the estimated panning angle, while no line means no activity. We observe that the overall trend is that the binary ASI resembles the activity of the sources, both in silent periods and when the sources contain significant energy.

The fundamental frequency estimates of the harmonic sources are obtained using the estimated panning parameters and the ASI. The mixture is downsampled to $f_s = 8$ kHz, and processed in segments of length $N = 480$ samples (60 ms), with a hop size of $H = 120$ samples (15 ms). The fundamental frequencies are estimated using a grid with 1 Hz spacing, from $f_{0,\text{min}} = 50$ Hz to $f_{0,\text{min}} = 1000$ Hz. As explained in Section 3.2, a codebook of magnitude amplitudes is used when estimating the complex amplitudes of the sources. The codebook is trained using anechoic instrument recordings from the IOWA database[1], and the signals

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

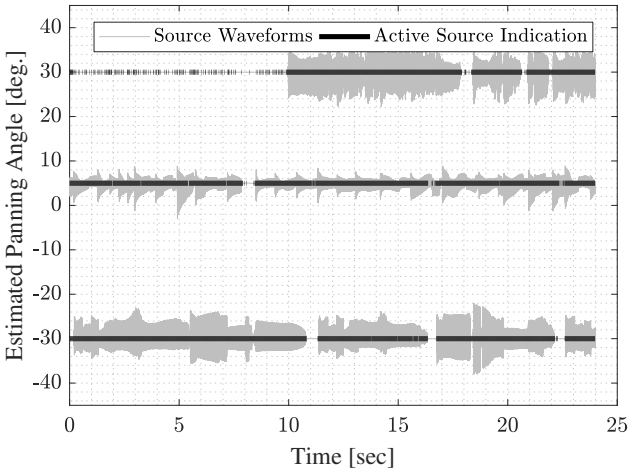

Figure 2: Active source indication (ASI) shown as black lines. For each frame of 15 ms there is an indicator. The ASI is overlayed on the original source signals which do not relate to the panning axis.

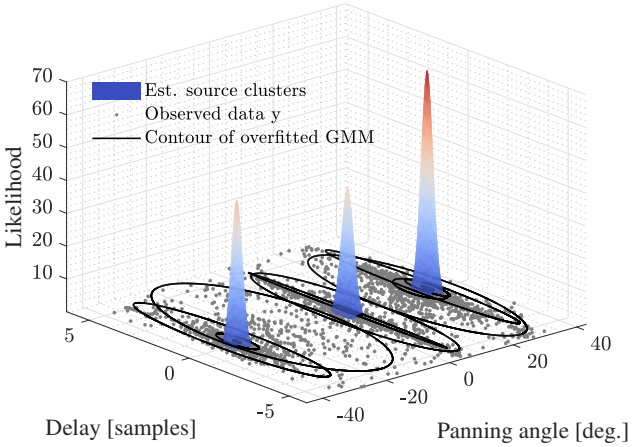

Figure 3: Proposed GMM estimation of source panning clusters.

used for training are listed in Table 3. See [17] for further details. The fundamental frequency estimates of the sources are shown in Figure 4, along with the ground truth which was obtained using the `joint_anls()` function from the Multi-Pitch Estimation Toolbox [24] on the individual instrument recordings from the dataset resulting in single-pitch estimates. No smoothing has been applied to the parameter estimates. The separation of the sources from the mixture is done using Wiener filtering, as described in Section 3.3. A spectrogram of a monophonic version of the observed mixture, obtained as an average of the stereo channels, is shown in Figure 6 along with the residual, which is obtained by subtracting the estimated sources from the mixture. We observe that most of the harmonic components in the mixture have been removed. The spectrograms of the unmixed and reconstructed bass tracks are shown in Figure 7. The reconstructed bass track contains most of the harmonic content in the unmixed source, however, some of the higher harmonics are missing. In Figure 8 the spectrograms of the unmixed and reconstructed trombone tracks are presented. The reconstructed trombone signal again contains most of the harmonic

Table 3: Data used for generating the amplitude codebook (v: played with vibrato).

Instrument	Instr. type	Note ranges	Duration (s)
Alto flute	Woodwind	G3-B3, C4-B4	68.3
Alto sax	Woodwind	Db3-B3, C4-B4	118.9
Alto sax (v)	Woodwind	Db3-B3, C4-B4	129.2
Bass flute	Woodwind	C3-B3, C4-B4	113.3
Bassoon	Woodwind	C3-B3, C4-B4	55.7
Bb clarinet	Woodwind	D3-B3, C4-B4	111.4
Eb clarinet	Woodwind	G3-B3,C4-B4	47.5
French horn	Brass	C2-B2, C4-B4	68.0
Oboe	Woodwind	Bb3-B3, C4-B4	46.6
Soprano sax	Woodwind	Ab3-B3, C4-B4	64.3
Soprano sax (v)	Woodwind	Ab3-B3, C4-B4	69.2
Tenor trombone	Brass	C3-B3, C4-B4	106.2
Trumpet	Brass	E3-B3, C4-B4	170.3
Trumpet (v)	Brass	E3-B3, C4-B4	182.9

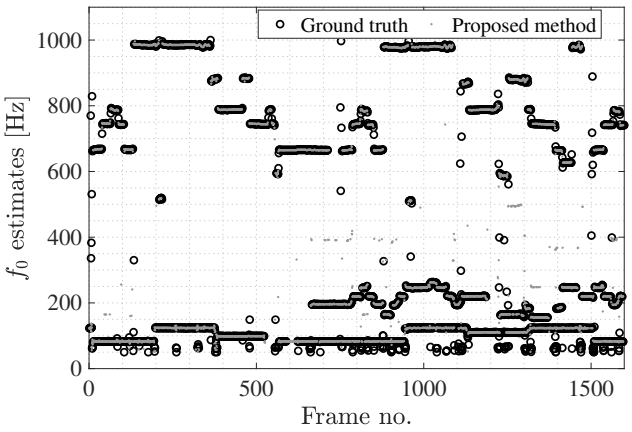

Figure 4: Fundamental frequency estimates of the sources in the mixture.

content, however, some segments in the beginning of the signal contain energy which was not present in the unmixed source; this is due to errors in the ASI. The spectrograms of the unmixed and reconstrcuted clarinet tracks are shown in Figure 9. Comparing the spectrograms of the unmixed and reconstructed tracks, it can be seen that the main harmonic components of the source have been captured in the reconstruction. A graphical user interface (GUI) is written in MATLAB in which the sources can be re-panned, using either the original panning parameters, or using new parameters[2]. Figure 5 shows a screenshot of the mixing GUI. An informal listening test suggests that including the residual ensures that information not captured by the harmonic model, such as breathing noises and other non-stationarities greatly improves the perceived quality of the reconstructed mixture.

[2]An audiovisual demonstration of the re-panning is available at `https://youtu.be/0HHoMVyOGcU`

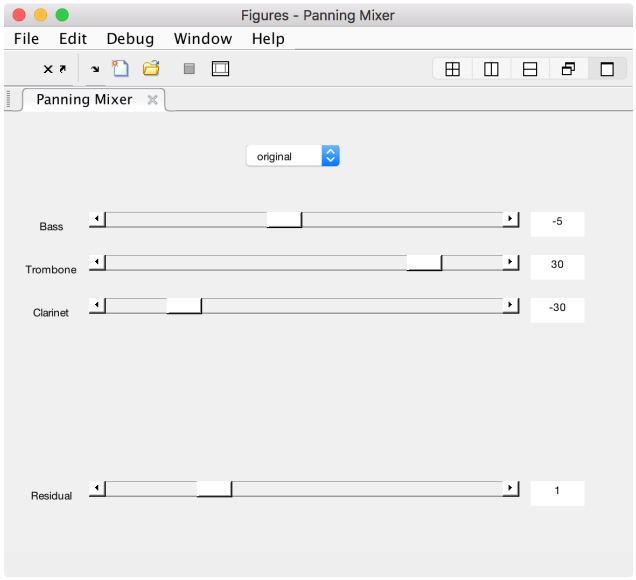

Figure 5: Screenshot of the GUI for mixture reconstruction.

5. DISCUSSION

In this paper, a method for separating an observed stereophonic mixture into its harmonic components, is presented. The method does not require knowledge of the number of sources in the mixture. The sources are extracted using a multi-channel harmonic signal model, where the panning parameters and the number of active sources in each frame of the mixture are estimated in an initial step. The fundamental frequencies, amplitudes and number of harmonics are estimated using an iterative approach. To enforce spectral smoothness, the magnitude amplitudes of the harmonics are mapped to entries in a codebook, which has been trained on individual notes played on a variation of instruments. The harmonic components are extracted by modelling the sources using the harmonic model and the estimated parameters. When the harmonic sources have been extracted, they are processed individually, i.e, the panning parameters of the sources are altered. The reconstruction of the mixture includes the residual, which contains the parts of the signal that are not captured by the harmonic signal model. When the residual is added to the mixture of extracted harmonic components, the resulting mixture is more pleasing to listen to. Extensions to this work could be the inclusion of inharmonicity in the signal model, to allow more precise modelling of string instrument signals, such as guitar, bass and piano recordings. Temporal smoothness could also be imposed in the parameter estimation steps. Furthermore, the signal model presented here is anechoic, i.e., the performance of the proposed method will degrade in the presense of reverberation effects. One option is to use a method for dereverberation, such as one of the methods presented in [31].

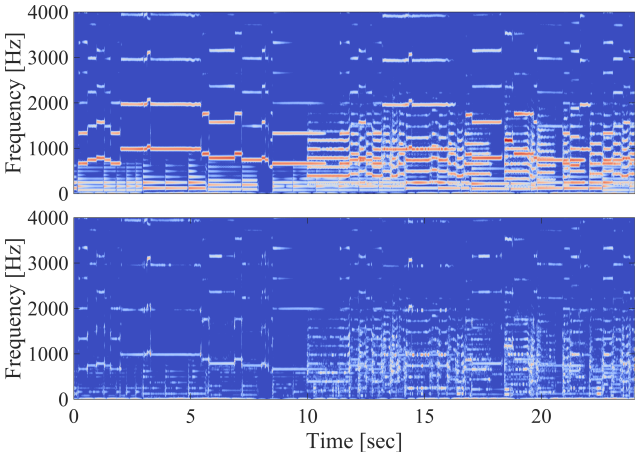

Figure 6: Spectrogram of the observed mixture (top) and the residual after subtraction of the harmonic sources (bottom).

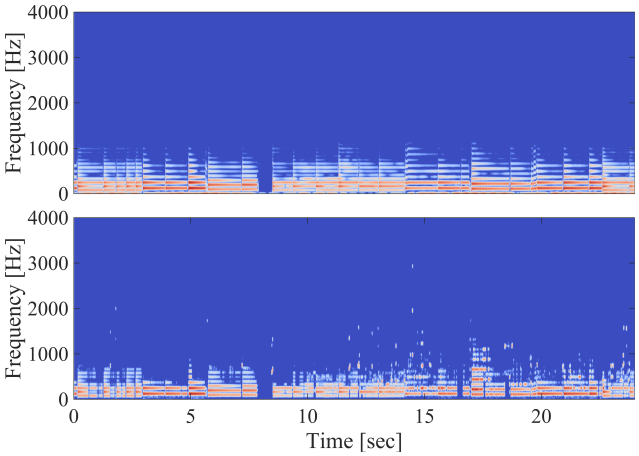

Figure 7: Spectrogram of the unmixed bass track (top) and the reconstructed bass track (bottom).

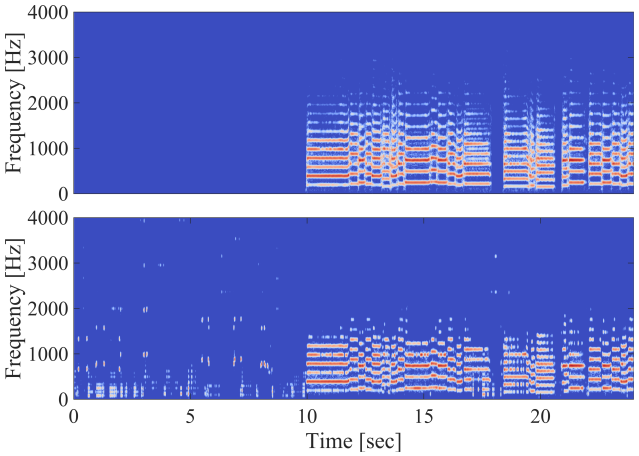

Figure 8: Spectrogram of the unmixed trombone track (top) and the reconstructed trombone track (bottom).

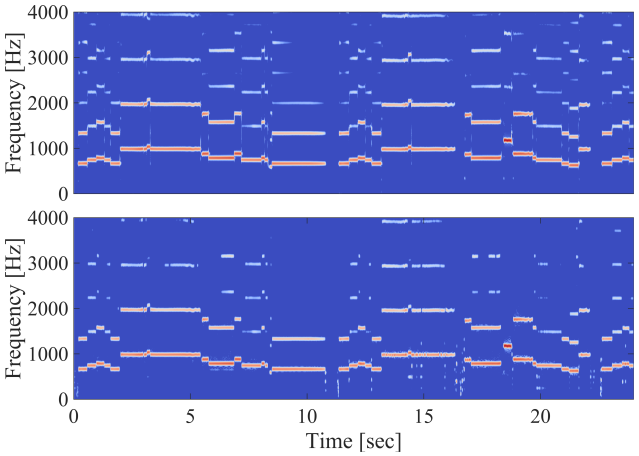

Figure 9: Spectrogram of the unmixed clarinet track (top) and the reconstructed clarinet track (bottom).

6. REFERENCES

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

[2] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Gado, V. Pulkki, and E. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 31–42, 2015.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, April 2017.

[4] L. Benaroya, L. M. Donagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, April 2003, vol. 6, pp. VI–613–16 vol.6.

[5] 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.

[6] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, March 2010.

[7] 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.

[8] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 281–295, Feb 2018.

[9] S. Leglaive, R. Badeau, and G. Richard, "Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, 2017.

[10] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct 2003, pp. 55–58.

[11] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999.

[12] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 2, pp. II765–II768 vol.2.

[13] J. Woodruff and B. Pardo, "Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings," *EURASIP J. on Applied Signal Processing*, vol. 2007, no. 1, pp. 086369, Dec 2006.

[14] J. Han and B. Pardo, "Reconstructing completely overlapped notes from musical mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011.

[15] A. Ben-Shalom and S. Dubnov, "Optimal filtering of an instrument sound in a mixed recording given approximate pitch prior," *Proc. Int. Computer Music Conf. (ICMC)*, 2004.

[16] 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.

[17] 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.

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

[19] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, vol. 1, pp. I–529–I–532.

[20] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 381–396, 2002.

[21] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, *Handbook of Cluster Analysis*, Chapman and Hall/CRC, 2015.

[22] 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.*, April 1993, vol. 2, pp. 728–731 vol.2.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research.," in *Proc. Int. Conf. Music Information Retrieval*, 2014, vol. 14, pp. 155–160.

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