SOUND SOURCE SEPARATION IN MONAURAL MUSIC SIGNALS USING EXCITATION-FILTER MODEL AND EM ALGORITHM

Anssi Klapuri

Queen Mary University of London,
Centre for Digital Music, London, UK

Tuomas Virtanen, Toni Heittola[*]

Tampere University of Technology,
Dept. of Signal Processing, Tampere, Finland

ABSTRACT

This paper proposes a method for separating the signals of individual musical instruments from monaural musical audio. The mixture signal is modeled as a sum of the spectra of individual musical sounds which are further represented as a product of excitations and filters. The excitations are restricted to harmonic spectra and their fundamental frequencies are estimated in advance using a multipitch estimator, whereas the filters are restricted to have smooth frequency responses by modeling them as a sum of elementary functions on Mel-frequency scale. A novel expectation-maximization (EM) algorithm is proposed which jointly learns the filter responses and organizes the excitations (musical notes) to filters (instruments). In simulations, the method achieved over 5 dB SNR improvement compared to the mixture signals when separating two or three musical instruments from each other. A slight further improvement was achieved by utilizing musical properties in the initialization of the algorithm.

***Index Terms*—** Sound source separation, excitation-filter model, maximum likelihood estimation, expectation maximization.

1. INTRODUCTION

Sound source separation means estimating the signals of individual sources from a mixture. The task is closely related to auditory scene analysis where, for humans, segregating the sounds of simultaneously active sources is an important part of making sense of complex auditory scenes. In this paper, we consider source separation in monaural music signals: separating the signals of individual musical instruments from a single-channel mixdown. Applications of this include musical instrument recognition in polyphonic audio, music remixing (emphasis or suppression of certain instruments), flexible processing and manipulation of music, audio coding, and analysis of the individual instruments' signals. Many of these applications do not require perfect separation quality, but robust segmentation of the time-frequency plane according to the sources.

Separation of multiple sources has been recently studied using various approaches. Some are based on grouping sinusoidal components to sources (see e.g. [1]) whereas some others utilize a structured signal model [2, 3]. Some methods are based on supervised learning of instrument-specific harmonic models [4], whereas recently several methods have been proposed based on unsupervised methods [5, 6, 7]. Some methods do not aim at separating time-domain signals, but extract the relevant information (such as instrument identities) directly in some other domain [8].

In this paper, we propose a source separation method based on the excitation-filter model of sound production. The excitation part corresponds to a vibrating system (such as a guitar string) and involves pitch information, whereas the filter corresponds to the body response of an instrument (such as the piano soundboard) which colours the spectrum of the sounds produced by the instrument. The excitation signals are first estimated using a multipitch estimator [9]. Then we propose a novel expectation-maximization (EM) algorithm which jointly learns the filter responses and organizes the excitations (notes) to filters (instruments). Contrary to the model we have proposed earlier in [10], here each note is assigned only to one instrument (with a certain probability), and the parameter estimation is done in the maximum likelihood sense. The proposed method is able to handle polyphonic sound sources that produce multiple notes simultaneously, for example separating piano and electric guitar from each other. As a side-product, the method produces note pitches and organization of notes to their instruments.

2. SIGNAL MODEL

2.1. Excitation-filter model

We use the excitation-filter signal model, where excitations correspond to different pitch values (notes) and these are filtered by the body response that is characteristic to each instrument. The resulting spectra of individual musical sounds are then summed to obtain the mixture magnitude spectrum $\hat{x}_t(k)$:

$$\hat{x}_t(k) = \sum_{n=1}^{N_t} g_{n,t} e_{n,t}(k) h_{i(n,t)}(k) \tag{1}$$

where $g_{n,t}$ is the gain of note n at time t and $e_{n,t}(k)$ is the note's excitation spectrum. Magnitude response of the filter corresponding to instrument i is denoted by $h_i(k)$ and $i(n,t)$ denotes the instrument that played note n at time t.

The filter $h_i(k)$ is further represented as a linear combination of fixed elementary responses

$$h_i(k) = \sum_{j=1}^{J} c_{i,j} a_j(k) \tag{2}$$

where we choose the elementary responses $a_j(k)$ to consist of triangular bandpass responses, uniformly distributed on the Mel-frequency scale $f_{\text{Mel}} = 2595 \log_{10}(1 + f_{\text{Hz}}/700)$. The responses are illustrated in Fig. 1. The J coefficients $c_{i,j}$ encode the spectral shape of instrument i. Representing the filters this way makes the estimation more robust since $J \ll K$ (number of frequency bins).

To start with, let us consider a situation where the note-instrument associations $i(n,t)$ are given. Estimating these will

[*] This work was supported by the Academy of Finland, (Centres of Excellence in Research 2006-2011)

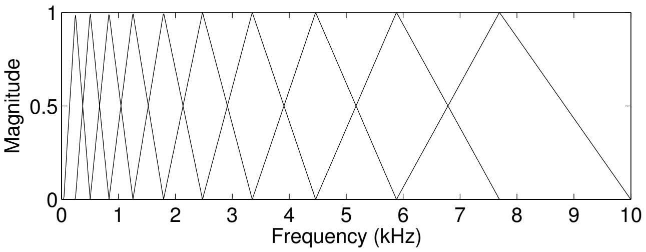

Fig. 1. The elementary responses used to represent the filters $h_i(k)$. The responses are uniformly distributed on the Mel-frequency scale.

be considered in the next subsection. For the sake of mathematical tractability, we assume the following observation noise model:

$$\mathsf{p}(x_t(k)|\theta, z_t) \propto \exp(-d(x_t(k), \hat{x}_t(k))) \tag{3}$$

where $\propto$ denotes equality up to a scalar multiplier that is independent of the model variables, and $d(x, \hat{x})$ denotes the divergence

$$d(x, \hat{x}) = x \log(x/\hat{x}) - x + \hat{x}. \tag{4}$$

In the likelihood function (3), $x_t(k)$ is the observed magnitude spectrum value and $\hat{x}_t(k)$ is the value given by the model (1)–(2). For convenience, we use $\theta \equiv \{g_{n,t}, \mathbf{e}_{n,t}, \mathbf{c}_i\}$ to denote all the model parameters at all times, and the symbol z_t to represent the information regarding all the note-instrument associations $i(n, t)$ at time t. The noise model (3) and maximum-likelihood estimation lead to minimizing KL divergence between the observations and the model, which has produced good results in earlier sound separation studies [7]. This is equivalent to assuming that the observations are generated by a Poisson process with mean value $\hat{x}_t(k)$.

2.2. Associating notes to instruments

Let us now consider the case where the note-instrument associations are not given in advance. If there are N_t concurrent notes at time t, there are I^{N_t} different ways of organizing the notes to I instruments. For example, if there are three notes and two different sound sources, the possible note-instrument associations are

$$\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}2\\1\\1\end{bmatrix}, \begin{bmatrix}1\\2\\1\end{bmatrix}, \begin{bmatrix}2\\2\\1\end{bmatrix}, \begin{bmatrix}1\\1\\2\end{bmatrix}, \begin{bmatrix}2\\1\\2\end{bmatrix}, \begin{bmatrix}1\\2\\2\end{bmatrix}, \begin{bmatrix}2\\2\\2\end{bmatrix}, \tag{5}$$

where the second vector, for example, means that note 1 is associated with instrument 2 and notes 2 and 3 are associated with instrument 1. The task of choosing one of the above vectors is here called *note labelling*: assigning each note a number which tells its sound source.

Let us arrange the different note-labelling alternatives in matrix $\mathbf{I}$ of size $(N \times I^N)$. Entries of the matrix are given by

$$i(n, z) = 1 + \bmod\left(\left\lfloor (z-1)/I^{(n-1)} \right\rfloor, I\right) \tag{6}$$

where $n = 1, \ldots, N$ and $z = 1, \ldots, I^N$. Here $\lfloor \cdot \rfloor$ denotes rounding towards the negative infinity and modulo $\bmod\,(x, y) \equiv x - y\lfloor x/y \rfloor$. It is easy to verify that in the case of three notes and two instruments, (6) will give a matrix $\mathbf{I}$ where the eight vectors (5) are as columns.

We will treat the integer z_t as an unknown latent variable that determines the instrument $i(n, z_t)$ of all notes n in the frame t. However, estimating z is here done only in a probabilistic sense. Let us denote by $\alpha_t(z) \equiv \mathsf{p}(z_t)$ the probability of note-labelling z at time t. In each frame t, the probabilities sum to unity, $\sum_z \alpha_t(z) = 1$. Note that the matrix $\mathbf{I}$ remains the same in all frames if the number of instruments I is constant: matrices for varying N_t are obtained as submatrix at the upper left corner of $\mathbf{I}$ calculated for maximum N_t.

In order to estimate the note-instrument associations z_t for $t = 1, \ldots, T$, we include the probabilities $\alpha_t(z)$ as a new parameter in our model. Let us denote the augmented parameter vector by $\Theta \equiv \{g_{n,t}, \mathbf{e}_{n,t}, \mathbf{c}_i, \alpha_t(z)\}$, where the probabilities $\alpha_t(z)$ for all z at all times are included.

The probability density function (pdf) of the observed spectrum $x_t(k)$ is now calculated by summing over the different note-instrument associations, weighted by their probabilities:

$$\mathsf{p}(x_t(k)|\Theta) \propto \sum_z \alpha_t(z) \exp\left[-d(x_t(k), \hat{x}_{t,z}(k)\right]. \tag{7}$$

Above, we have denoted

$$\begin{aligned}\hat{x}_{t,z}(k) &= \sum_{n=1}^{N_t} g_{n,t} e_{n,t}(k) h_{i(n,z)}(k) \\ &= \sum_{n=1}^{N_t} g_{n,t} e_{n,t}(k) \sum_{j=1}^{J} c_{i(n,z),j}\, a_j(k)\end{aligned} \tag{8}$$

where the latter form is obtained by substituting from (2). Please observe that above $i(n, z)$ is now the filter (instrument) index, varying along with z. It is worth comparing (7) with (3) where the note-instrument associations were assumed known.

We assume that the observation noise in all frames and at all frequencies is independent. This means that the observations in all the frames and frequencies are conditionally independent given the model parameters. Thus the whole pdf of $\mathbf{X}$ is given by

$$\begin{aligned}\mathsf{p}(\mathbf{X}|\Theta) &\propto \prod_{t,k} \sum_z \alpha_t(z) \exp\left[-d(x_t(k), \hat{x}_{t,z}(k)\right] \\ &= \prod_t \sum_z \alpha_t(z) \exp\left[-\sum_k d(x_t(k), \hat{x}_{t,z}(k)\right]\end{aligned} \tag{9}$$

The idea of the parameter estimation is to find such parameters Θ that the above likelihood function $\mathsf{p}(\mathbf{X}|\Theta)$ is maximized.

3. PARAMETER ESTIMATION

3.1. Multipitch analysis

The excitation spectra $e_{n,t}(k)$ are estimated independently of all the other model parameters. First, a multipitch estimator proposed by Klapuri in [9] is used to estimate note pitches $F_t(n), n = 1, \ldots, N_t$ in each analysis frame t.

Based on the pitch value $F_t(n)$, the corresponding excitation $e_{n,t}(k)$ is constructed by summing Hamming-windowed sinusoidal components at integer multiples of the pitch frequency $F_t(n)$. These “harmonic combs” extend over the entire frequency range considered (up to 10 kHz) and have a unity magnitude for all the harmonics. An example excitation spectrum is shown in Figure 2.

The number of concurrent notes (polyphony) N_t is not estimated at this stage, but instead, five simultaneous pitches are estimated in each frame. It is the task of the subsequent EM algorithm to find near-zero gains $g_{n,t}$ for the exraneous notes. In frames where the actual polyphony is higher than five, we assume that the estimated five pitches model the signal sufficiently well.

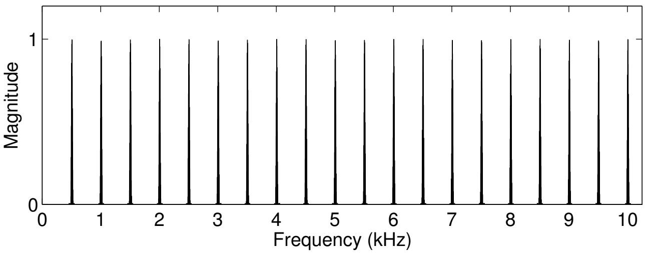

Fig. 2. An example excitation spectrum $e_{n,t}(k)$ corresponding to pitch value 500 Hz.

3.2. Expectation maximization algorithm

This section describes an iterative parameter estimation method based on the expectation-maximization (EM) algorithm [11]. EM algorithm operates by initializing the parameters Θ to some initial values Θ_0 and then iteratively repeating the E step and M step (described below), so as to increase the value of the likelihood function $\mathsf{p}(\mathbf{X}|\Theta)$ at each iteration until the likelihood value converges.

Let us denote by $\mathbf{z} = z_1 z_2 \cdots z_T$ the entire sequence of latent variables (note-instrument associations) over time. In the E step of the algorithm, we evaluate the posterior probabilities $\mathsf{p}(\mathbf{z}|\mathbf{X},\Theta)$ (in contrast with the priors $\alpha_t(z)$). The posterior probabilities $\beta_t(z) \equiv \mathsf{p}(z|\mathbf{X},\Theta)$ of different note-instrument associations z at time t are given by

$$\beta_t(z) = \frac{\alpha_t(z)\exp\left[-\sum_k d(x_t(k),\hat{x}_{t,z}(k))\right]}{\sum_{z'}\alpha_t(z')\exp\left[-\sum_k d(x_t(k),\hat{x}_{t,z'}(k))\right]} \tag{10}$$

where $\hat{x}_{t,z}(k)$ is defined in (8). The entire posterior $\mathsf{p}(\mathbf{z}|\mathbf{X},\Theta)$ is the product of marginals $\mathsf{p}(z|\mathbf{X},\Theta)$, since the observation noise is assumed independent in each frame.

In the M step, we calculate Θ^{new} given by

$$\Theta^{\text{new}} = \arg\max_{\Theta} Q(\Theta,\Theta^{\text{old}}) \tag{11}$$

where

$$\begin{aligned} Q(\Theta,\Theta^{\text{old}}) &= \mathsf{E}_{\mathbf{z}|\mathbf{X},\Theta}\left[\ln \mathsf{p}(\mathbf{X},\mathbf{z}|\Theta)\right] \\ &= \sum_{\mathbf{z}} \mathsf{p}(\mathbf{z}|\mathbf{X},\Theta)\ln \mathsf{p}(\mathbf{X},\mathbf{z}|\Theta). \end{aligned} \tag{12}$$

Above, summing over $\mathbf{z}$ means summing over all possible sequences $\mathbf{z} = z_1 z_2 \cdots z_T$. After some algebraic manipulation (omitted here for space constraints), the above formula simplifies to

$$Q(\Theta,\Theta^{\text{old}}) = \sum_t \sum_z \beta_t(z)\left[\ln \alpha_t(z) - d(x_t(k),\hat{x}_{t,z}(k))\right] \tag{13}$$

where the first term is maximized by updating $\alpha_t(z) \leftarrow \beta_t(z)$ and the latter term is recognized as weighted divergence. The entire function can be maximized by the following updates

$$\begin{aligned} g_{n,t} &\leftarrow g_{n,t}\frac{\sum_{i,j,k} e_{n,t}(k)c_{i,j}a_j(k)\sum_{z\in Z_{n,i}}\beta_t(z)\frac{x_t(k)}{\hat{x}_{t,z}(k)}}{\sum_{i,j,k} e_{n,t}(k)c_{i,j}a_j(k)\sum_{z\in Z_{n,i}}\beta_t(z)} \\ c_{i,j} &\leftarrow c_{i,j}\frac{\sum_{n,t,k} g_{n,t}e_{n,t}(k)a_j(k)\sum_{z\in Z_{n,i}}\beta_t(z)\frac{x_t(k)}{\hat{x}_{t,z}(k)}}{\sum_{n,t,k} g_{n,t}e_{n,t}(k)a_j(k)\sum_{z\in Z_{n,i}}\beta_t(z)} \\ \alpha_t(z) &\leftarrow \beta_t(z) \end{aligned} \tag{14}$$

where set $Z_{n,i}$ is defined as $Z_{n,i} = \{z : i(n,z) = i\}$. In other words, the summing in the update formulas is performed only over the values of z where note n is associated with instrument i.

Summary of the overall parameter estimation is the following:

1) Initialize the probabilities $\alpha_t(z)$ for all t and z with random noise uniformly distributed between zero and one. Normalize so that $\sum_z \alpha_t(z) = 1$ in all frames t. Initialize the gains $g_{n,t}$ and the filter coefficients $c_{i,j}$ with absolute values of Gaussian noise.
2) Evaluate $\beta_t(z)$ using (10).
3) Update gains $g_{n,t}$, filter coefficients $c_{i,j}$, and note-instrument association probabilities $\alpha_t(z)$ using (14). The updated parameters constitute Θ^{new}.
4) Evaluate the total likelihood $\ln \mathsf{p}(\mathbf{X}|\Theta)$ using (9) and return to Step 2 if the likelihood has changed more than some convergence threshold ϵ.

3.3. Resynthesis of source signals

Magnitude spectrograms corresponding to individual instruments i are estimated as

$$y_{i,t}(k) = \frac{\hat{x}^{\Theta}_{i,t}(k)}{\hat{x}^{\Theta}_{t}(k)}x_t(k) \tag{15}$$

where $x_t(k)$ is the observed mixture spectrum, $\hat{x}^{\Theta}_{i,t}(k)$ are instrument-specific spectra obtained from the model with parameters Θ,

$$\hat{x}^{\Theta}_{i,t}(k) = \sum_{n=1}^{N_t}\sum_{z\in Z_{n,i}}\alpha_t(z)g_{n,t}e_{n,t}(k)\sum_{j=1}^{J}c_{i(n,z),j}a_j(k) \tag{16}$$

and $x^{\Theta}_t(k) = \sum_i \hat{x}^{\Theta}_{i,t}(k)$.

Time-domain signals are generated by using phases of the mixture signal and inverse discrete Fourier transform.

3.4. Algorithm initialization utilizing musical properties

The above-presented estimation algorithm is already complete and able to separate the individual instruments' signals. In this section, we describe an "add-on feature" which utilizes musical assumptions of voice leading to improve over the random initialization of the algorithm in Step 1 of Sec. 3.2.

One property of voice leading in music is that consecutive notes arriving from the same singer/instrument are relatively close in pitch. Because the spectral shape (filter) of the instrument remains rather constant too, the entire spectra of notes arriving from the same instrument resemble each other.

The initialization of the proposed method can be slightly improved by tentatively grouping the excitations to instruments already after the multipitch estimation, based on their spectral shape. In practice, we computed Mel-frequency cepstral coefficients (MFCCs) of a spectrum that was constructed by picking only the spectral components corresponding to the harmonics of excitation n from the mixture spectrum $x_t(k)$. K-means clustering algorithm was then used to assign the excitations to I clusters. Then, instead of random initialization of $\alpha_t(z)$, we initialized them so that each excitation was given probability $1-(I-1)\eta$ of having come from the instrument to which it was tentatively assigned, and a smaller probability η of having come from the other instruments. Note that this only affects the initialization of the EM algorithm, which then updates and corrects the initial values of $\alpha_t(z)$.

This algorithm is denoted by "proposed method + musical initialization" in the simulations.

Table 1. Average SNRs of the separated signals

Number of instruments in mixture	2	3
SNR before separation	0.0 dB	-3.0 dB
Separated with proposed method	5.1 dB	2.4 dB
Proposed method + musical initialization	5.8 dB	2.6 dB

4. RESULTS

Simulation experiments were carried out to evaluate the performance of the proposed method. Acoustic signals for the experiments were obtained by synthesizing pieces from the RWC Popular Music database [12]. MIDI synthesis was used in order to obtain the signals of individual instruments which were then mixed to obtain the test signals. A total of 50 MIDI files were randomly selected from the database and then synthesized using Timidity software synthesizer and a high-quality GeneralUser GS 1.4 soundfont. To allow maximally realistic synthesis, all control messages in the MIDI files were retained.

A five-second segment was randomly selected from among the leading 90 seconds of each piece to constitute the test excerpt. The number of concurrent instruments was controlled so that that only the signals of two or three instruments were mixed to produce each test excerpt, in order to keep the separation task feasible. The instruments were randomly selected from among all pitched instruments in the piece, however so that the probability of including an instrument in the test case was proportional to the time span that the instrument was active in the randomly chosen segment. This was done to avoid including "silent most of the time" instruments.

It should be noted that although the number of instruments in a test file is only two or three, the number of concurrent notes can be much higher, since many instruments, such as piano or guitar, produce multiple notes at a time. The instruments occurring in our test signals were the following (occurrence frequencies in parentheses): a/e bass (31), a/e guitars (29; 19 of which electric), a/e pianos (25), flutes (22), strings (15), synthesizers (8), brass instruments (6), mallet percussions (4), organs (4), reed instruments (3), percussion/effects (3), where "a/e" denotes acoustic/electric.

Separation quality was measured by calculating the signal-to-noise ratio, SNR $= 10 \log_{10} \sum_t s(t)^2 / \sum_t [\hat{s}(t) - s(t)]^2$, where $s(t)$ and $\hat{s}(t)$ are the reference and the estimated signal, respectively. The SNRs are then averaged over all separated signals.

Table 1 shows the obtained average SNRs using the proposed source separation method. As can be seen, the proposed method achieves more than 5 dB SNR improvement compared to the mixture signal before separation. Including musical properties in the initialization bring an additional slight improvement. This verifies that the proposed method is able to separate sources from the mixture and organize the notes to the instruments in an unsupervised manner.

Measuring SNRs is not completely fair, since minimizing the divergence (4) does not minimize reconstruction error in least-squares sense, but emphasizes more small-magnitude spectral regions that are also perceptually more important. Examples of separated signals are available at http://www.cs.tut.fi/sgn/arg/klap/icassp2010/

5. CONCLUSIONS

A method was proposed for separating musical instruments from polyphonic music signals. The method is able to handle polyphonic instruments such as piano and guitar, and achieves over 5 dB SNR improvements (compared to the mixture signal before separation) without any prior information about the analyzed signals, except the number of sources I. This indicates that only assuming the general structure of music signals shown in the signal model (7)–(8) enables source separation in complex music signals. Source separation arises naturally from the estimation of the parameters of the structured signal model. As a side-product, the method produces note pitches and organization of notes to their instruments. Utilizing musical properties in the initialization of the algorithm yields a slight further improvement. A drawback of the method is relatively high computational complexity, especially for large I.

6. REFERENCES

[1] J.J. Burred, A. Röbel, and T. Sikora, "Dynamic spectral envelope modeling for the analysis of musical instrument sounds," *IEEE Trans. Audio, Speech, and Language Processing*, 2009.

[2] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proc. EUSIPCO*, Glasgow, Scotland, August 2009.

[3] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," in *Proc. IEEE ICASSP*, Taipei, Taiwan, 2009, pp. 3073–3076.

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

[5] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tenson factorisation models for musical source separation," *Computational Intelligence and Neuroscience*, 2008.

[6] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. IEEE ICASSP*, Las Vegas, USA, 2008.

[7] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[8] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrogram: Probabilistic representation of instrument existence for polyphonic music," *IPSJ Journal*, vol. 48, no. 1, pp. 214–226, 2007.

[9] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Intl. Conf. on Music Information Retrieval*, Victoria, Canada, 2006, pp. 216–221.

[10] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. 10th Intl. Society for Music Information Retrieval Conference*, Kobe, Japan, 2009, pp. 327–332.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[12] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC music database: Popular, classical, and jazz music databases," in *Intl. Conf. on Music Information Retrieval*, Paris, France, Oct. 2002, pp. 287–288.