Single-channel mixture decomposition using Bayesian harmonic models

Emmanuel Vincent, Mark Plumbley

Single-channel Mixture Decomposition using Bayesian Harmonic Models

Emmanuel Vincent and Mark D. Plumbley

Electronic Engineering Department, Queen Mary, University of London
Mile End Road, London E1 4NS, United Kingdom
emmanuel.vincent@elec.qmul.ac.uk

Abstract. We consider the source separation problem for single-channel music signals. After a brief review of existing methods, we focus on decomposing a mixture into components made of harmonic sinusoidal partials. We address this problem in the Bayesian framework by building a probabilistic model of the mixture combining generic priors for harmonicity, spectral envelope, note duration and continuity. Experiments suggest that the derived blind decomposition method leads to better separation results than nonnegative matrix factorization for certain mixtures.

1 Introduction

1.1 Constrained specific models and unconstrained generic models

Single-channel musical source separation is the problem of extracting the source signals $(s_j(t))_{1 \leq j \leq J}$ underlying a music signal $x(t) = \sum_{j=1}^{J} s_j(t)$. This problem can be addressed by building appropriate models of the sources. The source models proposed in the literature rely on different amounts of prior information.

Some methods exploit constrained source models representing the sources in a specific mixture with a good accuracy. For example, methods based on sparse coding with a fixed dictionary [1] or on factorial hidden Markov models [2] typically assume that the source models can be learnt on segments of the mixture where only one source is present. These methods provide very good separation results, given the difficulty of the problem, but until now they rely on knowing the instruments present in the mixture and performing a manual segmentation. Other methods based on Computational Auditory Scene Analysis (CASA) with instrument templates [3] or on hybrid source models [4] rely on instrument-specific timbre properties learnt on a database of isolated notes. These methods also perform satisfyingly, but they cannot be applied when some of the instruments present in the mixture are not part of the learning database.

By contrast, other methods rely on unconstrained generic source models applicable to a large range of mixtures. For example, Nonnegative Matrix Factorization (NMF) decomposes the mixture short-term magnitude spectrum into a sum of components modeled by a fixed magnitude spectrum and a time-varying gain, assuming no constraints about the spectra and the gains except positivity [5]. Source separation can then be achieved by clustering the components

into sources, provided each component belongs to a single source. Good results based on automatic clustering have been reported for the separation of vocals [6] or drums [7] from real mixtures. Other studies using a manual clustering have shown that NMF can be used to separate real mixtures of non-percussive instruments [8]. However the NMF source model is not adapted to certain types of mixtures, such as those involving notes with time-varying fundamental frequency, instruments with similar spectral envelope or instruments playing synchronously.

1.2 Harmonicity as a precise generic model

In this paper, we assume that each musical note is a near-periodic signal containing harmonic sinusoidal partials. Harmonicity means that at each instant the frequencies of the partials are multiples of a single fundamental frequency. This assumption is true for sustained instruments such as bowed strings and winds and approximately true for many other instruments. It is false for drums, human voice and other noisy or transient sounds. Harmonicity can thus be seen as a precise generic model: it gives more information about the sources than the NMF model while being valid for a large range of mixtures. In the following, we call *harmonic component* a set of harmonic partials having common onset and offset times and we address the problem of Harmonic Component Extraction (HCE), that is the decomposition of a mixture into such components. We do not discuss the difficult issue of clustering the estimated components into sources.

Most existing HCE methods consist in performing a polyphonic pitch tracking, that is transcribing the fundamental frequencies of the notes present in the mixture, and then estimating the amplitudes and phases of their harmonics. Methods exploiting harmonicity only [9] are insufficient for source separation. Indeed harmonicity does not provide enough information to segregate partials from different sources overlapping at the same frequency. Other methods have used complementary assumptions of spectral continuity [10,11] and temporal continuity [12,10] to this aim. Since polyphonic pitch tracking is a difficult problem for which no current algorithm provides a perfect solution, the separation performance of these methods was mostly evaluated based on prior knowledge of the fundamental frequencies and few quantitative results were reported.

In the following, we recast the problem of estimating harmonic components in the Bayesian framework. We model the mixture signal as a sum of harmonic components whose parameters are governed by probabilistic priors and we estimate the number of components and their parameters using a Maximum A Posteriori (MAP) criterion. This can be seen as a coherent approach where polyphonic pitch tracking and estimation of the amplitudes and phases of the partials are performed using the same model. The proposed model is inspired by Bayesian harmonic models introduced previously in the literature for polyphonic pitch transcription [13] but it includes several modifications. Most importantly, we design a perceptually motivated residual prior and we learn the parameters of other priors on a database of isolated notes rather than setting them manually to arbitrary values. When this learning database is large, the resulting model is generic. We have also used this model recently for object coding purposes [14].

The rest of the paper is structured as follows. Section 2 presents the generative model of the mixture and the associated inference algorithm. Section 3 compares the performance of the proposed method with NMF on a few test mixtures. Finally Section 4 discusses some future research directions.

2 Bayesian inference of harmonic components

2.1 Signal model

The proposed model is expressed in the time domain. Let $x_n(t)$ be the n-th frame of the mixture signal $x(t)$ defined by $x_n(t) = w(t)x(nS + t)$ where $w(t)$ is a Hanning window of length W and S is the stepsize. We develop $x_n(t)$ as

$$x_n(t) = \sum_{c \in \mathcal{C}_n} s_{cn}(t) + e_n(t), \tag{1}$$

where $(s_{cn}(t))_{c \in \mathcal{C}_n}$ are the harmonic components present in this frame and $e_n(t)$ is the residual. We define each harmonic component, which generally spans several time frames, by

$$s_{cn}(t) = w(t) \sum_{m=1}^{M_c} a_{cmn} \cos(2\pi m f_{cn} t + \phi_{cmn}), \tag{2}$$

where f_{cn} is its fundamental frequency and (a_{cmn},ϕ_{cmn}) are the time-varying amplitude and phase of its m-th partial in the n-th frame.

2.2 Frequency, amplitude and spectral envelope priors

We associate each component with a latent fundamental frequency F_c belonging to the MIDI scale, which is the discrete 1/12 octave scale used for western musical scores. We constrain the number of partials M_c of the c-th component to

$$M_c = \min((F_{\max}/F_c), M_{\max}), \tag{3}$$

where $F_{\max}$ is the Nyquist frequency and $M_{\max}$ is set to 60. On each time frame, we model the fundamental frequency by a log-Gaussian prior

$$P(\log f_{cn}) = \mathcal{N}(\log f_{cn}; \log F_c, \sigma^f), \tag{4}$$

where $\mathcal{N}(\cdot; \mu, \sigma)$ is the univariate Gaussian density of mean μ and standard deviation σ. In order to help estimate the amplitudes of the partials when partials from several notes overlap at the same frequency, we describe the amplitudes as the product of a fixed normalized spectral envelope $(\mu^a_{F_c m})_{1 \leq m \leq M_c}$, a latent log-Gaussian amplitude factor r_{cn} and a log-Gaussian residual, that is

$$P(\log a_{cmn} | r_{cn}) = \mathcal{N}(\log a_{cmn}; \log(r_{cn} \mu^a_{F_c m}), \sigma^a_{F_c}), \tag{5}$$

$$P(\log r_{cn}) = \mathcal{N}(\log r_{cn}; \mu^r_{F_c}, \sigma^r_{F_c}). \tag{6}$$

Finally we assume that the phases of the partials are uniformly distributed

$$P(\phi_{cmn}) = 1/(2\pi). \tag{7}$$

2.3 Duration and continuity priors

Perceptually annoying discontinuities may appear in the extracted source signals when the model parameters are estimated on each time frame separately. Thus we add duration and continuity priors on the parameters. We associate each point on the MIDI scale with a binary activity state in each frame determining whether a harmonic component with the corresponding latent frequency F_c is being played or not in that frame, with the constraint that different instruments cannot play notes with the same latent frequency at the same time. We assume that the sequences of activity states for different points on the MIDI scale are independent, and we model each sequence by a two-state Markov prior. We also set temporal continuity priors on the frequencies and amplitudes of the partials

$$P(\log f_{cn}|f_{c,n-1}) = \mathcal{N}(\log f_{cn}; \log f_{c,n-1}, \sigma^{f'}), \tag{8}$$

$$P(\log a_{cmn}|a_{cm,n-1}) = \mathcal{N}(\log a_{cmn}; \log a_{cm,n-1}, \sigma^{a'}_{F_c m}), \tag{9}$$

$$P(\log r_{cn}|r_{c,n-1}) = \mathcal{N}(\log r_{cn}; \log r_{c,n-1}, \sigma^{r'}_{F_c}). \tag{10}$$

The global prior on amplitudes and frequencies is then defined up to a multiplicative constant by multiplying these priors with the local priors defined above.

2.4 Perceptually motivated residual prior

The role of the prior on the residual is to ensure that the largest possible number of notes present in the mixture are extracted using a given number of components. The standard Gaussian prior measures the distortion between the mixture signal and the model according to the energy of the residual. This often results in several components being used to represent high-energy notes, while low energy parts of the mixture such as low energy notes, onsets and reverberation are not transcribed despite their perceptual significance. We design instead a weighted Gaussian prior inspired from the distortion measures proposed in [15,16] which give a larger weight to perceptually significant low energy parts.

The proposed prior models the first stages of auditory processing. The incoming sound first passes through the outer and the middle ear and is split by the cochlea into several frequency subbands called auditory bands. The energy in each auditory band is then transformed nonlinearly into a loudness value taking into account masking phenomena. More precisely, we measure the power of the residual in the b-th auditory band by $\tilde{E}_{nb} = \sum_{f=0}^{W/2} v_{bf} g_f |E_{nf}|^2$, where $(E_{nf})_{0 \leq f \leq W-1}$ are the Fourier transform coefficients of $e_n(t)$, $(v_{bf})_{0 \leq f \leq W/2}$ are coefficients modeling the frequency spread of that band and $(g_f)_{0 \leq f \leq W/2}$ is the frequency response of the outer and middle ear. We measure similarly the power of the mixture signal in that band by $\tilde{X}_{nb} = \sum_{f=0}^{W/2} v_{bf} g_f |X_{nf}|^2$. Then we define the distortion due to the residual on the n-th frame by $L_n = \sum_{b=1}^{B} \tilde{E}_{nb} \tilde{X}_{nb}^{-0.75}$. It can be shown that this distortion is approximately equal to the perceived loudness of the residual on that frame [16]. We derive the residual prior from the distortion by $P(e_n) \propto \exp(-L_n/(2\sigma^{e\,2}))$. This prior can also be expressed as

$$P(E_{nf}) = \mathcal{N}(E_{nf}; 0, \sigma^e \gamma_{nf}^{-1/2}) \tag{11}$$

where

$$\gamma_{nf} = \sum_{b=1}^{B} v_{bf} g_f \left(\sum_{f=0}^{W/2} v_{bf} g_f |X_{nf}|^2 \right)^{-0.75} . \quad (12)$$

2.5 Approximate inference of harmonic components

The signal model and the parameter priors define together a probabilistic generative model of the mixture signal that is used to infer the MAP values of the activity states and the frequency, amplitude and phase parameters representing a given mixture. Due to the complexity of the model, exact inference is intractable. We therefore use a three-step approximate inference procedure instead. First we estimate the MAP activity states and the corresponding MAP parameters on each time frame separately, then we refine the estimation of the states by adding the duration priors, and finally we refine the estimation of the parameters by keeping the states fixed and adding the continuity priors. More details about these steps are given in [14]. Each harmonic component is then directly synthesized from the corresponding parameters.

3 Evaluation

3.1 Training, performance measure and optimal clustering

We evaluate the proposed HCE method on test mixtures sampled at 22.05 kHz. Hyper-parameters of the generative model are set to the same values for all test mixtures: σ^f, $(\mu^a_{F_c m})$, $(\sigma^a_{F_c})$, $(\mu^r_{F_c})$, $(\sigma^r_{F_c})$, $\sigma^{f'}$, $(\sigma^{a'}_{F_c m})$ and $(\sigma^{r'}_{F_c})$ are learnt on part of the RWC[1] Musical Instrument Database whereas σ^e and the Markov transition probabilities are set manually. The frame parameters are set to $W = 1024$ (46 ms) and $S = 512$ (23 ms) and discrete fundamental frequencies span the range between MIDI 36 (65 Hz) and MIDI 100 (2640 Hz).

For comparison purposes, we also evaluate NMF on the same test mixtures. We write the NMF generative model as $|X_{nf}| = \sum_{c=1}^{C} p_{cf} q_{cn} + E_{nf}$, where $(p_{cf})_{0 \leq f \leq W/2}$ and $(q_{cn})_{0 \leq n \leq N-1}$ are the fixed spectrum and time-varying amplitude of the c-th nonnegative component respectively. We assume that these quantities are positive and that the residual E_{nf} follows the weighted Gaussian prior above. The total number of spectra C is fixed manually and the spectra and time-varying amplitudes are estimated using multiplicative update rules. Source signals including several spectra are then synthesized by inverse Fourier transform and overlap-add using the phase spectrum of the mixture signal. This algorithm is similar to the weighted NMF algorithm introduced in [16], except the definition of the time-frequency weights (γ_{nf}) is modified by taking into account overlap between auditory bands.

For evaluation purposes, we partition components produced by HCE or NMF into source clusters based on prior knowledge of the true sources. We define the

[1] `http://staff.aist.go.jp/m.goto/RWC-MDB/`

optimal clusters as those which maximize the overall source separation performance and we compute them using a beam search procedure. This "oracle" clustering is not feasible in realistic situations, however it allows the measurement of the best source separation quality potentially achievable.

The source separation performance is measured locally for each estimated source j around each time frame n using a local phase-blind Signal-to-Distortion Ratio (SDR) in decibels (dB) defined by

$$\mathrm{SDR}_{jn} = 10 \log_{10} \left(\frac{\sum_{l=0}^{W'-1} w'(l)^2 |S_{j,n+l,f}|^2}{\sum_{l=0}^{W'-1} w'(l)^2 (|\hat{S}_{j,n+l,f}| - |S_{j,n+l,f}|)^2} \right), \qquad (13)$$

where $w'(l)$ is a Hanning window of length $W' = 12$ frames and $(\hat{S}_{jnf})$ and (S_{jnf}) are the short-term Fourier transforms of the j-th estimated source and the j-th true source respectively. The overall performance is measured by a global SDR defined as the median of local SDRs for all sources and all time frames. We believe that this performance measure accounts better for subjective effects than the standard time-domain SDR. Indeed the ear is approximately phase-blind and the error perceived at a given time depends only on the power of the target signal at that time, not on its total energy. However the actual subjective performance is better assessed by listening to the estimated source signals.

3.2 Results

We consider two sets of test mixtures: ten mixtures of two sources using real sources from the SQAM database[2], and ten MIDI-synthesized mixtures from the RWC Classical Music and Music Genre Databases containing two to five sources. We set the number of nonnegative components of NMF to be the same as the number of harmonic components estimated by HCE. This allows a rather fair comparison of the two methods, since in a blind context the difficulty of component clustering would depend on the number of components. We also separate MIDI-synthesized mixtures by HCE using knowledge of the note activity states. All the mixture signals and some of the estimated source signals are available for listening on `http://www.elec.qmul.ac.uk/people/emmanuelv/ICA06/`.

Table 1 shows that the global SDR achieved by HCE is on average 3 dB higher than NMF on mixtures of real sources and 6 dB higher on MIDI-synthesized mixtures. Informal listening tests suggest that the estimation errors made by the two methods are very different. As expected, NMF often fails to separate synchronized notes in MIDI-synthesized mixtures because these notes have the same temporal evolution. This results in strong interference or in continuous artifacts. More surprisingly, NMF also produces artifacts on mixtures of real sources which are not synchronized. By contrast, HCE generally produces fewer artifacts, but some interference appears locally due to simultaneous or successive notes with the same frequency being fused into a single component, or to harmonic partials from different sources being transcribed as part of the same component.

[2] `http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/`

Table 1. Comparison of the separation performance achieved by HCE and NMF.

Separation method	Global SDR on various mixtures of real sources (dB)									
HCE	12.9	13.3	13.8	10.9	19.3	17.3	14.6	15.2	14.7	11.9
NMF	10.3	11.6	12.6	7.2	14.0	11.8	10.9	11.9	13.0	10.4
Separation method	Global SDR on various MIDI-synthesized mixtures (dB)									
HCE with true score	14.5	29.3	10.8	13.0	10.4	3.0	12.3	9.4	17.7	8.8
HCE	15.4	29.3	7.2	11.9	10.7	5.5	11.5	3.2	17.0	5.1
NMF	6.9	7.4	5.7	7.0	1.4	3.5	3.4	2.6	14.5	3.3

The knowledge of the note activity states does not substantially improve the performance of HCE for seven out of ten MIDI-synthesized mixtures[3]. It is interesting to note that the number of notes estimated by HCE on MIDI-synthesized mixtures is on average 2.5 times larger than the actual number of notes being played. Most of the spurious notes have short duration and are due to the system trying to represent non-harmonic parts of the signal using harmonic components, which does not seem to affect the separation performance.

Other experiments suggested that the performance of NMF decreases when more components are allowed and does not change significantly when initializing the NMF basis spectra by the spectra of the harmonic components estimated by HCE. Thus the limited performance of NMF on the test mixtures seems to be the effect of the model itself rather than algorithmic issues.

4 Conclusion

In this paper, we address the blind source separation problem for single-channel musical mixtures where the notes are near-periodic signals containing harmonic sinusoidal partials. The proposed method, which exploits harmonicity and other generic source priors, performs better than NMF on various test mixtures. This suggests that the NMF model is not sufficiently constrained to ensure that typical audio source properties hold for the separated sources and that more precise generic source models can help separation without needing specific information about a particular mixture.

The main limitation of HCE is that it cannot deal with mixtures containing voice or drum instruments. This limitation could be addressed using a three-component generative model including probabilistic models for wideband noise components and transient components, in the spirit of the CASA system proposed in [12]. The proposed model could also be improved by adding slightly inharmonic components to represent instruments such as piano or guitar or by performing automatic adaptation of the probabilistic priors to the mixture to increase their precision and help reduce separation errors.

[3] For some mixtures the estimated note activity states lead to a better SDR than the true states because the perceptual weights used for decomposition are not taken into account for evaluation. In practice, the subjective performance of HCE using the true note activity states is always larger or equal to that of blind HCE.

References

1. Benaroya, L., McDonagh, L., Bimbot, F., Gribonval, R.: Non negative sparse representation for Wiener based source separation with a single sensor. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). (2004) VI–613–616
2. Ozerov, A., Philippe, P., Gribonval, R., Bimbot, F.: One microphone singing voice separation using source-adapted models. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). (2005) 90–93
3. Kinoshita, T., Sakai, S., Tanaka, H.: Musical sound source identification based on frequency component adaptation. In: Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI) Workshop on Computational Auditory Scene Analysis. (1999) 18–24
4. Vincent, E.: Musical source separation using time-frequency source priors. IEEE Trans. on Speech and Audio Processing **14(1)** (2006) To appear.
5. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). (2003) 177–180
6. Vembu, S., Baumann, S.: Separation of vocals from polyphonic audio recordings. In: Proc. Int. Conf. on Music Information Retrieval (ISMIR). (2005) 337–344
7. Helén, M., Virtanen, T.: Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In: Proc. European Signal Processing Conf. (EUSIPCO). (2005)
8. Wang, B., Plumbley, M.D.: Musical audio stream separation by non-negative matrix factorization. In: Proc. UK Digital Music Research Network (DMRN) Summer Conf. (2005)
9. Gribonval, R., Bacry, E.: Harmonic decomposition of audio signals with matching pursuit. IEEE Trans. on Signal Processing **51** (2003) 101–111
10. Virtanen, T.: Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint. In: Proc. Int. Conf. on Digital Audio Effects (DAFx). (2003) 35–40
11. Every, M.R., Szymanski, J.E.: Separation of synchronous pitched notes by spectral filtering of harmonics. In: IEEE Trans. on Speech and Audio Processing. (2006) To appear.
12. Ellis, D.P.W.: Prediction-driven computational auditory scene analysis. PhD thesis, Dept. of Electrical Engineering and Computer Science, MIT (1996)
13. Davy, M., Godsill, S.: Bayesian harmonic models for musical pitch estimation and analysis. Technical Report CUED/F-INFENG/TR.431, Cambridge University (2002)
14. Vincent, E., Plumbley, M.D.: A prototype system for object coding of musical audio. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). (2005) 239–242
15. van de Par, S., Kohlrausch, A., Charestan, G., Heusdens, R.: A new psychoacoustical masking model for audio coding applications. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). (2002) II–1805–1808
16. Virtanen, T.: Separation of sound sources by convolutive sparse coding. In: Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA). (2004)