

Methods of Single-Channel Music Source Separation

Matthew J. Crossley

June 13, 2010

Introduction

Music source separation refers to the process of recovering original music sources from a mixture of two or more musical sound sources. Although music source separation is important even when the number of mixture channels is high (e.g. in stereo or surround sound mixtures), this review is focused on music source separation when the number of mixtures is limited to a single-channel. This task can be formalized as follows: Assume that a single-channel audio signal x is created by summing M individual sources s_m .

$$x(n) = \sum_{m=1}^M s_m(n), \quad n = 1, \dots, N \quad (1)$$

Here, N is the length of the audio signal. In the context of equation 1, music source separation refers to the task of recovering each s_m given only the mixture x . A variety of methods have been used for single-channel music source separation to varying degrees of success. This review describes the basics of these methods, and provides a simple conceptual framework in which they can be related to each other.

Music source separation is closely related to Blind Source Separation (BSS). As such, it isn't too surprising that most attempts at single-channel music source separation have used techniques that are largely derived from the BSS field. In the context of music source separation, however, these techniques have taken on unique characteristics. This essentially because In true BSS, nothing is known about the sources or the mixing process. In contrast, we often know (or can easily infer) useful characteristics about a mixture of musical sounds. Single-channel music source separation systems differ in the amount of knowledge they assume or infer about the sources and mixing process. As we will see, this forms the basis some of the main differences between the major algorithms.

All music source separation methods have a fairly simple, and remarkable common layout. First, all systems must choose some representation of the music signal. Next, they must choose a mixture-model that they assume generated the single-channel mixture of musical signals. The observed data (the single-channel mixture) is then used in some way to derive estimates of the parameters of the mixture-model. These parameter estimates are

then used to generate estimated source signals. Methods differ primarily in the model used to describe the supposed mixing process, and the cost-function used to derive parameter estimates to the model.

The main models used to describe the mixing process are the linear instantaneous mixing model, the convolutive source model, the instrument models, and models based on probabilistic latent variables. The linear instantaneous mixing model is the source model used in Independent Component Analysis (ICA) and most Nonnegative Matrix Factorization (NMF) systems. In this model, the sources are assumed to add linearly and instantaneously to form the observed mixture. The convolutive source model assumes that the sources are convolved with each other to form the observed mixture. Instrument models make strong assumptions about the time-frequency structure of the sources, but in so doing make much weaker assumptions about the mixing process than either of the two models previously discussed. Finally, models based on probabilistic latent variables assume that the mixture process is probabilistic. These methods require training with at least some of the original sources prior to the source estimation process, but are very flexible with respect to the types of mixtures they can resolve.

There is relatively little variance in the type of representation most single-channel music source separation systems use for the observed mixture signal. This is because most of the mixing models just described require the number of observed mixture-signals to be greater than or equal to the number of source-signals to be estimated. Single-channel separation procedures that use these methods must then find some way of generating many channels of mixture data from a single source. This is most commonly done by generating a spectrogram (or long series of spectrograms as is usually the case) from the observed mixture-signal. These spectrograms are typically generated by taking the Short-Time Fourier Transform (STFT) of a set of windowed, partially overlapping chunks of the original mixture-signal. However, these spectrograms can, in principle, be generated with any invertible transform. In fact, recent work has suggested that Lapped Orthogonal Transforms (LOTs), which use Discrete Cosine Transforms (DCT) outperform methods based on the STFT in single-channel music separation.

The final major area that source-separation systems differ is how they estimate the parameters of their mixing-models. In general, all systems must first decide on original parameter estimates (which can be and is often a totally random guess), generate estimated source-signals, assess the quality of the estimation, and update the mixing-model parameter estimates until the estimated sources meet some system-defined criterion. This is typically done by constructing a cost-function that is a function of the parameter values of the mixing-model, and finding a parameter set that minimizes this cost-function. The systems we review have used cost-functions designed such that estimated sources have a variety of important properties such as statistical independence, nonnegativity, sparseness, and temporal continuity. Of course, measures of reconstruction error (i.e., some function of the difference between the observed mixture-signal and the estimated mixture-signal) are often used in the construction of the cost-function.

The remainder of this review describes each of these modeling attributes in more detail. Section 2 describes the various mixture-process models, section 3 describes the various mixture-signal models, and section 4 describes the various cost-functions used in the most widely known single-channel music source separation systems.

Mixture-Process Models

Linear Instantaneous Mixing Model

The simplest model of the mixture process is the the linear instantaneous mixing model. The linear instantaneous linear mixing model assumes that the sources are added linearly and instantaneously to form the observed mixture-signal. Formally, for the t^{th} frame of a mixture-signal $x_t(n)$

$$x_t(n) = \sum_{j=1}^J g_{j,t} b_j(n), \quad t = 1, \dots, T \quad n = 1, \dots, N/T \quad (2)$$

where J is the number of basis functions, and $g_{j,t}$ is the gain of the j^{th} basis function in the t^{th} frame. Since the same equation is used for every frame of the mixture-signal and we can conveniently combine equations from each frame into a single matrix equation as

$$\mathbf{X} = \mathbf{B}\mathbf{G} \quad (3)$$

where \mathbf{X} , \mathbf{B} , and \mathbf{G} are matrices. Each column of \mathbf{X} is one frame of the mixture signal x , each column of \mathbf{B} is the corresponding basis function, and each row of \mathbf{G} is the corresponding time-varying gain. In the context of equation 3, the goal of single-channel music source separation is to estimate \mathbf{B} , and \mathbf{G} given only the observed mixture \mathbf{X} .

Convolutional Source Model

The convolutional source model models the mixture-signal in each frame as the convolution between a source-signal event and a function that characterizes the onsets and gains of that event. This is formally expressed as,

$$x_t(n) = \sum_{\tau=1}^D a_n(t - \tau) s_{n,f}(\tau) \quad (4)$$

where $s_{n,f}(\tau)$ is a single event of source n at discrete frequency f , D is the duration of the longest of the $s_{n,f}(\tau)$ events, and $a_n(t)$ is the function that characterizes the onset and gain of each event.

Probabilistic Latent Variables

The probabilistic latent variable model is,

$$P(f, t) = \sum_z P(z)P(f|z)P(t|z) \quad (5)$$

where $P(f, t)$ is the spectrogram of the mixture-signal, and $P(f|z)$ and $P(t|z)$ are the prior distributions of the frequency-domain representation of the mixture-signal and the time-domain representation of the mixture-signal, respectively, conditioned on the latent variable z .

Mixture-Signal Representations

The Short Time Fourier Transform (STFT) of the input signal is by far the most common input signal representation used in the literature. This is probably because of the familiarity to many other signal processing applications. The STFT of the log STFT of the input mixture-signal is called the cepstrum and is a signal representation that is common in many other music information retrieval tasks. Interestingly, we know of no music source separation systems that use the cepstrum as the input mixture-signal representation. However, some methods have used simple time-domain signal representations and more recently some systems have used Lapped Orthogonal Transforms (LOTs) for the input mixture-signal representation. LOTs convert the input signal using basis functions that are essentially windowed discrete cosine functions, and as such are closely related to the discrete cosine transform (DCT).

Parameter Estimation and Cost-Functions

Music source separation is achieved in every mixture-process model we previously discussed (i.e., equations 2, 4, and 5) by choosing an input signal representation and using it somehow to estimate the parameters of the model (i.e., the terms on the right-hand sides of equation 2, 4, and 5). The criteria by which these parameters are estimated dramatically influence the results of resulting source estimates, and by far represent the greatest variance between music source separation systems. The basic idea behind all methods of parameter estimation is to simply initialize all parameters randomly, compute estimated sources using the model, and then evaluate the goodness of that model using some cost function. We will next describe several different criteria that have been used to create cost functions for music source separation.

Similarity Measures

Perhaps the most obvious and straight-forward method for estimating parameters of a model is to simply try to maximize the similarity between the estimated mixture and the observed mixture. There are many ways of computing similarity from multidimensional data such as mixtures of music source signals. Common examples of similarity measures that have been used by music source separation systems are the squared error (which forms the basis of the common statistical method sum of squared errors, or SSE) and the divergence. Even a modest review of such methods is beyond the scope of this paper, as there are already several great reviews of these methods in print.

Independence

One method of estimating model parameters is to make the recovered source as statistically independent as possible. When this method is employed within the linear mixing model the resulting system is equivalent to the popular blind source separation algorithm independent component analyses (ICA). Many music source separation systems have used ICA. One major concern with ICA is that, although statistical independence may be a reasonable assumption for many types of sources, it isn't a very good assumption for music sources. This is because music sources are very often highly statistically dependent as they often synchronously vary in time and pitch elements.

Nonnegativity

When working within a STFT- or LOT-based signal representation, the mixtures and sources are all nonnegative (i.e., every observed value is zero or positive). It turns out that this assumption alone is adequate for simple source separation tasks. When nonnegativity is used as the cost function within the linear mixing model the resulting method is called nonnegative matrix factorization (NMF). Like ICA, NMF has been used by several music source separation systems, and, as we have discussed, the assumption nonnegativity is a good assumption for music source signals.

Sparseness

In music source separation systems that use the STFT as the input mixture-signal representation, each frequency bin of the resulting spectrogram is typically treated as a single mixture. It is unlikely that any given source will have concentrated energy within more than a few frequency bins at a single time. This property of music signals is exploited by cost functions that include a sparseness term. The assumption of t sparsity manifests itself within ICA or NMF as an additional constraint on the time-varying gains (i.e., $g_{j,t}$ in equation 2 and $a_n(t)$ in equation 4). Specifically, nonzero $g_{j,t}$ are heavily penalized by the cost function.

Temporal Continuity

Temporal continuity refers to the idea that most musical signals vary slowly relative to the frame rate. As such, sudden large changes in the input signal are unlikely. The assumption of temporal continuity to estimate sources, like the assumption of sparsity, manifests itself within ICA or NMF as an additional constraint on the time-varying gains (i.e., $g_{j,t}$ in equation 2 and $a_n(t)$ in equation 4). Specifically, large differences between $g_{j,t}$ and $g_{j,t+1}$ are heavily penalized by the cost function.

Source Grouping and Classification

As a closing section, it is important to point out that when using the models in equations 2, 4 and 5, the recovered sources do not necessarily completely represent an underlying music source. Hence, it is typically necessary to classify and group the resulting sources such that each resulting group of signals represents an underlying source. The methods used to accomplish this grouping are beyond the scope of this paper, however, we wish to emphasize their importance in creating a function music source separation system.

References

- [1] Abdallah, S. A., and Plumbley, M. D., *Polyphonic Music Transcription by Nonnegative Sparse Coding of Power Spectra*
- [2] Barry, D., Fitzgerald, D., Coyle, E., and Lawlor, B., *Drum Source Separation using Percussive Feature Detection and Special Modulation*
- [3] Fitzgerald, D., Cranitch, M., and Coyle, E., *Non-Negative Tensor Factorization for Sound Source Separation*
- [4] Fitzgerald, D., Cranitch, M., and Coyle, E., *Extended Nonnegative Tensor Factorization Models for Musical Sound Source Separation*
- [5] Helen, M., Virtanen, T., *Separation of Drums From Polyphonic Music Using Non-Negative Matrix Factorization and Support Vector Machine*
- [6] Jafari, M. G., Abdallah, S. A., Plumbley, M. D., and Davies, M. E., (2006) *Sparse Coding For Convolutional Blind Audio Source Separation*
- [7] Nesbit, A., Vincent, E., and Plumbley, M. D., *Extension of Sparse, Adaptive Signal Decomposition to Semi-Blind Audio Source Separation*
- [8] Nesbit, A., Vincent, E., and Plumbley, M. D., *Benchmarking Flexible Adaptive Time-Frequency Transforms for Undetermined Audio Source Separation*

- [9] Paulus, J., Virtanen , T. *Drum Transcription with Non-Negative Spectrogram Factorization*
- [10] Plumbley, M. D., *Sparse Representations in Audio and Music: from Coding to Source Separation*
- [11] Shashanka. M., Raj, B., and Smaragdis, P. *Probabilistic Latent Variable Models as Non-Negative Factorizations*
- [12] Smaragdis, P., Shashanka. M., and Raj, B. *A Sparse Non-Parametric Approach for Single Channel Separation of Known Sounds*
- [13] Smaragdis, P. (2004) *Non-Negative Matrix Factor Deconvolution: Extraction of Multiple Sound Sources from Monophonic Inputs*
- [14] Smaragdis, P., Raj, B., and Shashanka. M., *Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures*
- [15] Smaragdis, P. and Mysore, G. J., *Separation by "Humming": User-Guided Sound Source Extraction From Monophonic Mixtures*
- [16] Smaragdis, P. and Brown, J. C., *Non-Negative Matrix Factorization for Polyphonic Music Transcription*
- [17] Virtanen , T. (2006) *Sound Source Separation in Monaural Music Signals*, Tampereen University of Technology. Publication 626.
- [18] Virtanen , T. *Separation of Sound Sources by Convolutional Sparse Coding*
- [19] Wang, B. and Plumbley, M. D., *Musical Audio Stream Separation by Non-Negative Matrix Factorization*