SEPARATING SOURCES FROM SINGLE-CHANNEL MUSICAL MATERIAL: A REVIEW AND FUTURE DIRECTIONS

Georgios Siamantas
Audio Lab
Department of Electronics
University of York

Mark R. Every
Centre for Vision,
Speech and Signal Processing
University of Surrey

John E. Szymanski
Audio Lab
Department of Electronics
University of York

ABSTRACT

The problem of separating multiple audio streams from single-channel polyphonic source material is a very challenging one, since it is extremely underdetermined. Additional information is hence essential to assist with constraining the infinite solution set, depending on the application. This paper reviews the projects in our group that have addressed this problem along with related research done elsewhere, illustrating how the possible future approaches for the current doctorial study follow naturally from this foundation.
Keywords – Source separation, mono, musical signal analysis

1. INTRODUCTION

Research in audio source separation has been growing steadily over the last 15 years or so. It is usual for audio signals coming from different sources to exist simultaneously in the form of a mixture. In Western music, this mixture is most often a set of interweaving melodies coming from some selection of pitched instruments, usually along with rhythmic percussive events coming from unpitched or pitched inharmonic instruments. The observation of this mixture can be done through a number of different sensors, or channels. Many existing separation methods have been exploiting these different multiple observations to extract the required signals without using any other form of information. This is called the Blind Audio Source Separation (BASS) problem. An extreme case of this problem is when there is only one available observation of the signal mixture (i.e. one channel). This individual case is considered an important problem by itself, since its solution could be useful in a wide variety of applications. Depending on the type of separation sought and the accuracy and fidelity which an algorithm might be able to deliver, possible applications include:

- The separation of the single-track source into multiple tracks, each corresponding to different instruments - effectively "demixing" the source - allowing an enhanced reprocessing of monophonic material, as well as its conversion to true stereo.
- The extraction of just a single source (instrument, voice) from the single-track source, offering the potential for adding new orchestrations to archive material.
- The separation not just of individual notes corresponding to a given musical source, but also the segmentation of different portions (e.g. onset and sustain) of each note, allowing individual creative control over each component part of the source.
- Enhanced access to the audio for restoration, repair, denoising and forensic purposes.
- The automatic transcription of musical material.
- Improved approaches to automatic classification of musical material for archiving, marketing and internet delivery.

Whereas some of these applications require extremely high fidelity of the resultant audio, others may employ the separation only as a means to an end, and hence reliable statistical measures of temporal variations in timing, pitch and amplitude for all instruments in the mix may be rather more important than the perceived audio quality. Further, a number of applications could be benefited by a real-time system, while some specific tasks, such as voice extraction from recordings, involve signals with particularly challenging characteristics.

In all cases, the fundamental difficulty with the separation of a single source recording into multiple streams is that the problem is extremely underdetermined [1] . It is an example of an ill-conditioned "inverse problem" where there are an infinite number of possible solutions. Additional information (that is often associated with inevitable assumptions) is hence essential to assist with constraining the solution set and rejecting unacceptable solutions, depending on the application.

There is a wide variety of approaches to this problem from different research areas of signal processing and computing [2] . The following section gives a brief overview of some of the approaches towards separating audio sources from single-channel mixtures focusing on their implied assumptions, models and current limitations.

2. REVIEW OF APPROACHES

2.1. The single-channel source separation problem

The task of retrieving the audio signals that correspond to individual sources from a single-channel signal can be formulated as the procedure of extracting the signals $s_j(t)$ from the mixture signal $x(t) = \sum_{j=1}^{J} s_j(t)$ where J is the number of sources. Because of the use of various assumptions and models, this procedure can be termed as a *semi-blind* source separation problem,

Email: gs510@ohm.york.ac.uk

[1] More accurately, it is the extreme situation of the underdetermined case of the general BASS or BSS (Blind Signal Separation) problem, where the number of channels is equal to the number of sources.

[2] Here we mention only the technical/engineering side of the problem. There are other important areas of research, though, such as music psychology, cognitive science and musical acoustics, whose contribution have been invaluable in providing crucial knowledge and insights.

as opposed to the general class of the BSS problems [3].
Due to the reasons mentioned in the previous section and the extreme indeterminacy of the 'one-to-many' problem, the various approaches differ in many aspects. Hence, it is important to stress also some of their basic assumptions, models and major restrictions.

2.2. Assumptions, models and limitations

Because of the fact that the audio signal under question is 'musical', much of the work has been done using a simplified (but rather useful for these early stages) definition of what is musical:

- Audio coming mostly from pitched and unpitched common acoustic instruments.
- The musical pitched sounds (i.e. the melody) are structured in discrete notes with relatively constant pitch and well defined onsets and offsets (what is loosely called a 'mid-level' audio signal structure [1]).
- The notes are modelled using the harmonic model (appearing in the frequency domain as a series of slowly-time varying frequency components comprised of the fundamental and the overtones placed at integer multiples of the fundamental's frequency) or near-harmonic model.
- The notes are modelled as containing a deterministic component (harmonic model) and an inharmonic component (transient or impulsive structure and shaped noise).

We continue, now, with some of the other assumptions and implicit limitations which are often encountered in previous work:

- Statistical independence of the sources.
- W-disjoint orthogonality (sources are assumed to be non-overlapping at each time-freq cell).
- Inherent limitations of the time-frequency representations (trade-off between time and frequency resolution [10]).
- Type and number of sources (e.g. only voice, strings, only percussion).

For the case of musical signals, statistical independence disregards the fact that the source signals often 'depend' on each other: For example, instruments playing at the same tempo, or playing harmonically related melodies (e.g. at the same key). Furthermore, for the musical case and for pitched instruments W-disjoint orthogonality means non-overlapping partials. This also implies the additional restriction that harmonically related notes (note intervals such as octaves, fifths and thirds) or, even worse, shared notes between instruments are not allowed. This is a major restriction for musical signals, since harmonic intervals appear more often in general than dissonant ones. The assumptions of statistical independence and W-disjoint orthogonality are very common among approaches that use spectral decomposition methods, like sparse coding [14] and Nonnegative Matrix Factorization (NMF) [22, 8]. In all this work no use of any music-inspired constraints, like harmonicity, is made. Recently, though, in [19] the harmonic model was indeed considered (in a bayesian framework), leading to a better performance than NMF on various mixtures. These methods, although leading to relatively good separation results, they are still limited from the fact that they cannot deal with overlapping partials of harmonic or near-harmonic sounds.

[3] As mentioned in [18], "[BSS] algorithms are generally expected to use as few assumptions as possible".

What's more, other methods like the ones based on factorial hidden Markov models [15, 13] rely on source models trained beforehand or by solo segments in the mixture, making them inapplicable if the source-specific models are not known.
Statistical independency, W-disjoint orthogonality or training procedures are not used by a number of other approaches. Instead, they use musical source waveform models, such as the sinusoidal harmonic model [21, 17]. Others use this model along with additional assumptions of spectral and temporal continuity (e.g. [20]). The separation of harmonic sounds is done by using an additive synthesis approach: The sinusoidal components are subtracted from the spectrum and then used to generate the original source signals using additive synthesis. These methods propose also techniques for dealing with overlapping frequency components.
Finally, it has to be noted that the final step of the separation process, which is the *automatic* clustering of the extracted components into different sources, is not considered by the majority of the above approaches (except in [8]).

3. RESEARCH AT YORK

There have been a number of recent projects at York which have investigated the use of model-based information, physical and musical constraints, instrument-specific data and *a priori* user-inserted information to assist with the separation of multiple audio streams from single-channel sources. The work has concentrated on non-real-time approaches and a wide range of musical signals, including some sung vocals, but is not designed for speech purposes.
In [12] the use of wavelet processing with Linear Predictive Coding (LPC) is employed in an attempt to separate the deterministic and stochastic part, as an alternative to Spectral Modelling Synthesis (SMS) [16]. It was found that for non-harmonic sounds such as percussive instruments/drums, the proposed technique gave more natural sounding results than SMS.
During an early attempt at single-channel source separation [9] an emphasis was given to how different viewpoints of the spectral information can contribute towards a better identification of harmonic structures. In particular, a Bayesian Belief Network (BBN) was introduced, where different points of view of the short-time spectrum (such as its magnitude, its autocorrelation and the spectrum of a slightly shifted analysis window) contribute to a decision whether a frequency component is part of a certain harmonic structure or not. In that case, these harmonic structures were used to identify the notes [4]. Although the model for the notes is obviously a simplification, some satisfying results were achieved for the separation of two singing voices.
Two other source separation attempts approached the problem in a different way: Identifying (manually) and estimating a "spectral fingerprint" of sections of background noise and then using that to subtract this noise from the mixture [3]; and extracting note harmonics in an iterative way [11] (discussed below).
In [7] an adaptive spectral filtering methodology is employed in conjunction with a near-harmonic model for extracting individual notes, exploiting a user-inserted MIDI score (rough note pitch and timings) as *a priori* information. This filtering approach was proposed as an alternative to the use of sinusoidal modelling for overcoming its shortcomings in source separation: The sinusoidal model is too strict to be able to estimate with high accuracy the non-stationary parameters of musical signals.

[4] More specifically, the General Linear Harmonic Model (GLHM) is used [9].

The method offers also a solution to the problem of overlapping spectral peaks in a single-channel recording. The overall separation results achieved using the adaptive filtering methodology in [5] were better than those where the nonlinear least-squares method [17] and linear equation solutions [2] were used to estimate the parameters of overlapping partials. Nevertheless, at segments with low SNR or with rapidly changing or poorly modelled signal content the filtering process cannot resolve the problem of interfering background noise or unmodelled signal content (for example when percussive sounds are leaking into the filtered notes).

After this filtering process, though, the unmodelled inharmonic content of the notes remains (as expected) in the residual. This content, that is mainly representative of the note onsets, plays an important role in the timbre perception and the naturalness of a sound. So, recognising the importance of incorporating appropriately the inharmonic content in the separation procedure, a three-step attempt was suggested towards this goal [4]: using an autoregressive model for decomposing transient and non-transient content; a bandwise noise interpolation technique for separating overlapping transient content [6]; and the use of correlation between harmonic and inharmonic content in individual notes for separating overlapping noise content. However, it was decided that it was too early to integrate these last three methods into the whole separation process, because they were considered not properly refined and compared with other methods [5].

Finally, a basic mechanism for automated grouping of the extracted notes into sources was implemented, allowing also the estimation of the number of sources. For the grouping of notes an unsupervised clustering algorithm was used. It was found that the algorithm would give better separation results using the original note segments than the extracted ones [5].

4. PROPOSED RESEARCH DIRECTIONS

4.1. A fully automatic separation system

Our primary goal is a system of separating source signals from polyphonic musical mixtures, the target applications of which will not necessarily require extremely high fidelity and precision for the resulting audio. This system will have to be a *fully automatic* system. Ideally this means that it will use a minimum number of assumptions and amount of prior information so that it can operate well enough with a wide range of mixtures. For the time being, though, this is considered unrealisable due to the current limitations discussed in the previous sections. Nevertheless, the "least possible" assumptions will be used.

Our work will initially built upon the general framework of the approach used in [5]. The first reason for choosing this is the fact that a relatively small amount of (quite common) assumptions is implied:

- Musical signal consisted of notes.
- Each note is consisted of near-harmonic or structured partials, transient and noise content.
- Use of STFT as the main tool for representing time-frequency structure.

Secondly:

- It does not assume statistical independence between sources.
- There are no inherent restrictions for the number and types of sources.
- It can deal with overlapping harmonic and inharmonic content.
- No prior learning is required.
- It proposes an automatic (unsupervised) method for clustering the notes into sources.

Finally, the main prior information that has been exploited in this framework, the MIDI score, will be replaced by a system that will give pitch and note timing information automatically. We are aware of the fact that the task of Automatic Music Transcription (AMT) is still an unsolved problem. However, we believe that it is a good time for a fully automatic single-channel source separation system to exist and to be tried. And the current framework in [5] seems like a good candidate, for the reasons discussed above. Moreover, there is the potential that the transcription process will be benefited by the tasks of iterative filtering of harmonics and other methods outlined below.

4.2. An iterative approach to spectral filtering

With regard to the extraction of harmonic content, the filtering algorithm that has been suggested previously is a one-pass approach: For each time segment it detects and extracts the spectral peaks corresponding to *all* the notes at once. As far as the 'simple' mixtures of pitched sounds are considered, this approach has been proven to be quite adequate. However, for real recordings, where a variety of pitched and unpitched instruments coexist, the performance has to be better. One way to go forward is by using an *iterative* subtraction scheme. As it was shown in [11] (using a rather simpler overall approach) this scheme can potentially lead to better sounding results. It can be implemented for each note segment as follows: The harmonics corresponding to the most prominent note are detected and extracted first. Then, the algorithm goes back to the residual and detects the next prominent group of harmonics. This process is repeated until all the notes have been extracted. In parallel with this process though, the residual will be re-examined each time to search for remnants of partials potentially belonging to the previously extracted notes and that the detection algorithm may have missed on previous iterations. The hypothesis that this alternative could lead to better separation of harmonics is based on the fact that the accuracy of the detection and estimation of the harmonics (especially of the weak ones) will increase, simply because there will be less interference after each extraction stage. It has to be noted though, that in order for this scheme to work as expected it has to be ensured that the extraction method does not introduce artifacts in the residual that could cause an error propagation between consecutive iterations.

4.3. Sinusoidal subtraction vs. spectral filtering

We are also going to address the inability of the current approach to stop percussive interference leaking into the filtered harmonic content. As pointed out in [5], employing the sinusoidal model to reconstruct the harmonics could be beneficial in cases where the noise level is considerably high. We will explore this possibility by modifying the extracting process to switch from filtering to sinusoidal subtraction, when the SNR is sufficiently low and there are no overlapping harmonics.

4.4. Extraction of inharmonic content

As mentioned above, the inharmonic part of the notes (transient and noise content) remains in the residual after subtraction of harmonic content. However, since the automatic classifier in [5] performed better using the original note segments than the

filtered ones, it is reasonable to believe that the information remaining in the residual could enhance its performance [5]. So, we intend to focus a part of our work on refining and possibly extending the capability of the methods mentioned in [4].

4.5. Other future considerations

Towards the improvement of the algorithm in identifying correctly the note harmonics, the BBN approach proposed in [9] (properly modified for sounds containing inharmonic components) will also be employed.
Finally, the system's performance can be tested in a number of possible situations, such as some Music Information Retrieval (MIR) tasks. One example of these would be to compare the performance of a score-matching task with/without using note grouping/source clustering.

5. CONCLUSION

In this paper, we addressed the challenging problem of single-channel source separation from polyphonic music recordings. Its solution can lead to a variety of applications, ranging from AMT and MIR to creative musical applications. After a brief review of methods carried out in our group and elsewhere with explicit reference to their assumptions, models and limitations, we presented how our future work can be shaped towards a fully automatic source separation system.

6. REFERENCES

[1] Bregman A. S. *Auditory Scene Analysis, The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts, 1990.

[2] Depalle P. and Hélie T. "Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows". In *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'97)*, New Paltz, NY, Oct. 19-22 1997.

[3] Dickin J. "Audio signal separation". Master's thesis, Department of Electronics, University of York, U.K., June 2004.

[4] Every M. R. "Separating harmonic and inharmonic note content from real mono recordings". In *Proc. Digital Music Research Network Summer Conf. 2005*, pp. 9–13, Glasgow, U.K., July 23-24 2005.

[5] Every M. R. *Separation of musical sources and structure from single-channel polyphonic recordings*. PhD thesis, Department of Electronics, University of York, U.K., 2006.

[6] Every M. R. and Szymanski J. E. "Separation of overlapping impulsive sounds by bandwise noise interpolation". In *Proc. 8th Int. Conf. on Digital Audio Effects (DAFx'05)*, Madrid, Spain, Sep. 20-22 2005.

[7] Every M. R. and Szymanski J. E. "Separation of synchronous pitched notes by spectral filtering of harmonics". *to be published in IEEE Trans. Audio, Speech and Language Processing*, 2006.

[5] Clearly, there are also other issues to be considered regarding the clustering process, such as trying different feature sets and alternative unsupervised clustering algorithms. We believe, though, that the inclusion of transient and noise information to the separation problem is of higher priority.

[8] Helén M. and Virtanen T. "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine". In *13th European Signal Processing Conference*, Antalaya, 2005.

[9] Jones J. G. L. "A transformation and segregation approach for the application of an interface driven by musical signals". DPhil Transfer Report, Feb. 2000.

[10] Masri P., Bateman A., and Canagarajah N. "A review of time-frequency representations, with application to sound/music analysis-resynthesis". *Organised Sound*, 2(3):193–205, 1997.

[11] McMillan J. "Mono to stereo and beyond". Master's thesis, Department of Electronics, University of York, U.K., June 2004.

[12] Mergen P. "Investigation into the use of wavelets for transformations of complex time-varying sounds". Master's thesis, University of York, Departments of Electronics and Music, U.K., Sept. 2001.

[13] Ozerov A., Philippe P., Gribonval R., and Bimbor F. "One microphone singing voice separation using source-adapted models". *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*, pp. 90–93, 2005.

[14] Plumbley M. D., Abdallah S. A., Bello J. P., Davies M. E., Monti G., and Sandler M. B. "Automatic music transcription and audio source separation". *Cybernetics and Systems*, 33(6):603–627, 2002.

[15] Roweis S. "One microphone source separation". *Advances in Neural Information Processing Systems (NIPS 13)*, pp. 793–799, 2001.

[16] Serra X. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

[17] Tolonen T. "Methods for separation of harmonic sound sources using sinusoidal modeling". In *presented at AES 106th Convention*, Munich, Germany, May 8-11 1999.

[18] Vincent E., Jafari M. G., Abdallah S. A., Plumbley M. D., and Davies M. E. "Blind audio source separation". Technical report, Centre for Digital Music, Queen Mary University of London, 24 November 2005.

[19] Vincent E. and Plumbley M. D. "Single-channel mixture decomposition using bayesian harmonic models". In *Proc. of the 6th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA 2006)*, number LNCS 3889, pp. 722–730, Charleston, SC, USA, 5-8 March 2006. Springer-Verlag, Berlin.

[20] Virtanen T. "Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint". In *Proc. Int. Conf. on Digital Audio Effects (DAFx).(2003)*, pp. 35–40, 2003.

[21] Virtanen T. and Klapuri A. "Separation of harmonic sound sources using sinusoidal modeling". In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'00)*, pp. 765–769, Istanbul, Turkey, June 2000.

[22] Wang B. and Plumbley M. D. "Musical audio stream separation by non-negative matrix factorization". In *Proc. Digital Music Research Network Summer Conf. 2005*, Glasgow, U.K., 23-24 July 2005.