Score-Guided Music Audio Source Separation

Christopher Raphael
School of Informatics
Indiana University
craphael@indiana.edu

Audio source separation seeks to decompose an audio recording into several different recordings corresponding to independent sources, such as speakers, foreground and foreground, or, in our case, musical parts. Source separation is a formidable task; while the problem has received considerable attention in recent years, it remains completely open.

The majority of approaches we know of are deemed *blind* source separation, meaning that the audio is decomposed without explicit knowledge of its contents [1] [2], [3]. In particular, much recent work has focused on Independent Component Analysis (ICA) [4] [5], as the methodological backbone of various approaches. Work on blind separation also contains work specifically devoted to music audio, such as [6] and [7]. While blind separation is, no doubt, broadly useful and deeply interesting, many of the techniques rely on restrictive assumptions about the recording process or audio, often not satisfied in practice. Moreover, blind approaches seem simply wrong-headed for our purposes, since they fail to capitalize on our explicit and detailed knowledge of the audio. The focus of our effort here is in fully incorporating this knowledge in a principled approach to musical source separation.

Our motivation stems from our ongoing work in musical accompaniment systems, in which a computer program generates a flexible and responsive accompaniment to a live soloist in a non-improvisatory piece of music. Our favorite musical domain is the *concerto*, or other work involving an entire orchestra for the accompaniment. Since our preferred approach resynthesizes a preexisting audio recording to synchronize with the live player [8], we rely on *orchestra-only* recordings. Some orchestral accompaniments can be purchased from commercial sources, however, the small collection of available accompaniments tend to be poorly recorded with variable playing. The ability to *desolo* a complete recording would open up a vast library of beautifully played and expertly recorded accompaniments for our system. Thus, our particular vantage point produces an asymmetrical view of the source separation problem, in which we seek to separate a single instrument from a large ensemble. This has important implications for the types of models and algorithms that we employ.

The unusual aspect of our problem statement is that we assume detailed knowledge of the audio content of our recordings. We begin with a symbolic musical score, giving the complete collection of pitches and rhythms in the solo and all accompanying parts. Additionally, our long-standing interest in score alignment has led to algorithms that automatically create a correspondence between the audio recordings and the symbolic scores [9], [10]. Thus, at any moment in the audio we know what notes are sounding and which parts they belong to. A partial depiction of our score knowledge is given in Figure in which vertical lines mark the onsets of each solo note. A similar problem statement was defined in [11].

While our interest is motivated by a particular application, this work potentially has broader impact. The most obvious application is *karaoke*, which also requires an accompaniment-

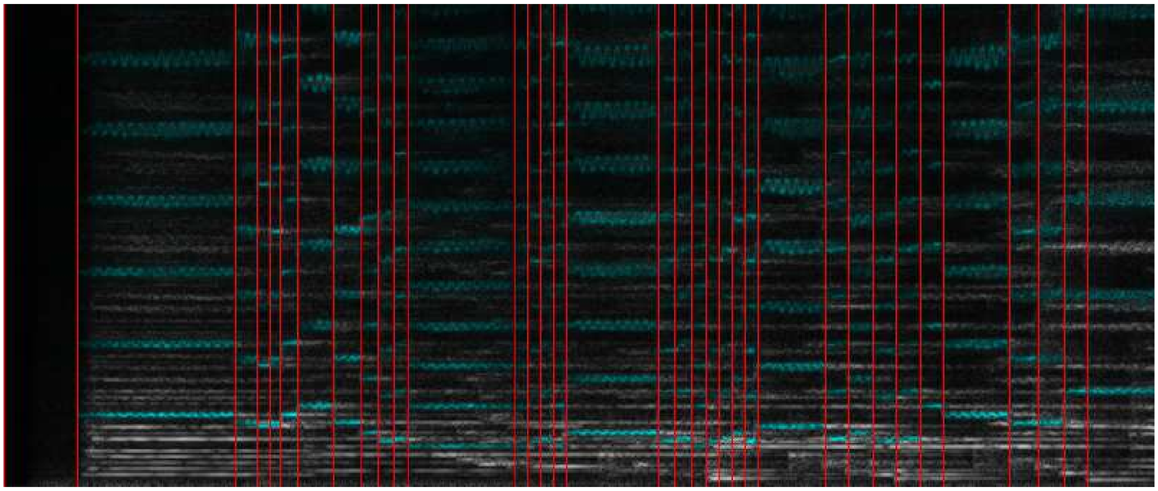

Figure 1: Spectrogram of opening of Samuel Barber Violin Concerto. The solo part is highlighted in blue while the solo note onsets are marked with red vertical lines.

only recording. Desoloing a recording is easy when the solo part is recorded separately and asymmetrically mixed into stereo channels, as is often the case in popular music: one need only estimate the mixing weights for each channel and then invert mixing operation. This popular technique, formalized by the most vanilla-flavored version of ICA, forms the basis of several commercial desoloing software products. When the recording and mixing techniques do not support this "trick," then methods such as our current proposal constitute a viable alternative. Other applications of the general problem of musical source separation include remixing existing recordings, incorporating existing musical material into new compositions, construction of audio databases, audio editing, and, no doubt, many ideas not yet conceived.

Our essential approach examines a small, yet reasonable, subset of possible decompositions of the audio: using our road map, we attribute each short-time Fourier transform (STFT) time-frequency point to either the soloist or to the accompaniment. Then, we invert our STFT using the appropriate subset of points to produce either the desoloed audio, or the soloist alone. This is the idea of "masking" that is now well-established in the source-separation community.

The benefit of phrasing the problem in terms of masking is that one only need estimate a boolean value for each STFT point: Does it belong to the soloist or orchestra? Surely this simplified problem is much easier than seeking an unconstrained decomposition into components that sum to the observed signal. The downside of masking is, of course, that only a small subset of possible decomposition can be expressed as masked versions of the original STFT. One might imagine that this leads to an inevitable degradation in the quality of the reconstructed audio. However, in synthetic experiments with "pre-mixed" data, we have demonstrated that the ideal mask leads to quite good audio reconstruction. Thus, the loss in audio quality that masking produces may well be compensated by the simplification of the problem statement. An example of the results of applying an ideal mask with premixed data are given at http://xavier.informatics.indiana.edu/˜craphael/nips06.

Our basic approach is to identify two types of solo events (subsets of STFT points) we seek to identify: solo note harmonics and transient events. Using the score match, the location of each solo note harmonic is approximately known, so we essentially seek a kind of "pitch-track" in which the *width* of the pitch interval also may vary with time. We have formulated this problem as seeking to identify the most likely labeling of STFT points using a probability model that was learned on pre-mixed data, in which the solo points are constrained to form a connected region in STFT space. The optimal labeling can then be found using dynamic programming.

We have formulated the identification of the transients in a similar way, though in this case search over the *vertical* extent of the STFT. Examples of our results on a real recording of the Samuel Barber Violin Concerto can be heard on the above-mentioned web page. From inspection of the STFT, this is clearly a very difficult example.

Even with our precise road map to the audio, our desoloing process degrades the resulting audio. While we hope to improve on our results, we expect this will always be true. In an unusual turn of events, however, forces seem to conspire in our favor to ameliorate this situation in the context of our accompaniment system. The damage done to the audio will be at the precise points in time-frequency space where the *live* soloist will be playing, thus masking much of the harm done in removing the recorded soloist. The web page shows an example of our accompaniment system using desoloed audio on the 2nd movement of the Strauss Oboe Concerto with the author playing the oboe. The desoloing procedure was more simple-minded than that presented here, but still produces good results.

The most significant contribution of this work is the recognition that a matched score can serve as the basis for musical source separation. This technique is generally applicable, in the sense that it does not rely on unrealistic assumptions about the recording process.

While we believe in posing the separation problem as one of estimating binary masks, there are many other, perhaps better, ways this estimation might be accomplished. The matched score can serve as the basis for estimating more detailed models of the signal, including the functions $|X_s|$ and $|X_a|$, or even the complete complex X_s and X_a, where X_s and X_a denote the solo and orchestral contributions to the STFT. The first of these, however, is complicated by the fact that $|X_s| + |X_a| \neq |X|$ as well as the difficulty imposed by the positivity restriction on our estimates, though this latter issue is an active research area [12]. When dealing with the full complex STFTs we *do* have $X_s + X_a = X$, however, it is unclear to us how to model the complex evolution of the signal. Both of these approaches are reasonable endeavors, even if the eventual goal is only the binary masks, since the extra nuisance parameters may lead the more precise estimation of the masks. Members of the Bayesian Signal Analysis community, as well as others, may recognize these as problems "right down their alley." We welcome the contributions of such areas and will endeavor to make score-matched audio data available to those who request it.

References

[1] Bregman, A., (1990) "Auditory Scene Analysis," MIT Press, 1990.

[2] Cardoso, J., (1998) "Blind signal separation: statistical principles,"" *Proceedings of the IEEE, special issue on blind identification and estimation,* vol. 9, no. 10, pp. 2009–2025, 1998.

[3] (1996) Ellis, D., "Prediction-driven computational auditory scene analysis,"" Ph.D. Dissertation, MIT Department of Electrical Engineering and Computer Science, 1996.

[4] Lee, T. W., Girolami, M., Bell A., Sejnowski, T. J. (1999) "A Unifying Information-theoretic Framework for Independent Component Analysis" *Int. journal of computers and mathematics with applications*, 1999.

[5] Bell, A. J., and Sejnowski, T. J., (1995) "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[6] Maher, R. C. (1990) "Evaluation of a Method for Separating Digitized Duet Signals" *J. Audio Eng. Soc.*, vol. 38, no. 12, pp. 956–979, 1990.

[7] Vincent, E., "Musical Source Separation Using Time-Frequency Source Priors," to appear in *IEEE Transactions on Speech and Audio Processing*
Neural Computation, vol. 7, no. 6, pp. 1129–1159, 1995.

[8] Raphael, C. (2003) "Orchestral Musical Accompaniment from Synthesized Audio," *Proceedings of the International Computer Music Conference* Singapore, 2003.

[9] Raphael, C. (1999) "Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models", *IEEE Trans. on PAMI* vol. 21, no. 4, 1999.

[10] Raphael, C., (2004) "A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores," *Proceedings of the 5th International Conference on Music Information Retrieval,* Ed. Claudia Lomeli Buyoli and Ramon Louriero, Barcelona, Spain, 387–394, 2004.

[11] Ben-Shalom, A., Shalev-Shwartz, S., Werman, M., Dubnov, S. (2004) "Optimal Filtering of an Instrument Sound in a Mixed Recording Using Harmonic Model and Score Alignment," *Proceedings of the ICMC*, 2004.

[12] Lee D. D., Seung S., (2000) "Algorithms for Non-negative Matrix Factorization" *Neural Information Processing Systems*, pp. 556–562, 2000.