A Classifier-Based Approch to Score-Guided Music Audio Source Separation

Christopher Raphael*

June 26, 2007

Abstract

We present an approach to the "desoloing" problem in which we seek to separate a soloist from the accompanying instruments using monaural audio. Our approach uses a time-aligned symbolic musical score; thus the musical pitches we wish to eliminate, as well as their time localizations are known. Separation is achieved by masking the short time Fourier transform (STFT). Thus the problem is simplified to classifying each STFT point as belonging to solo or accompaniment. Our experiments concern the problem of separating a violin soloist from a full orchestra. We first treat the problem with real audio data in which the contributions from soloist and orchestra are known. In this case we label time-frequency points using a classifier, learned from labeled training data, whose accuracy can be measured. We then extend these results to incorporate realistic constraints on the labeling. This latter method is tested on data from a commercial compact disc.

1 Introduction

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

Many approaches this audio decomposition problem 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

*this work supported by NSF grant IIS-0534694

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 symbolic musical score, giving the complete collection of pitches and rhythms in the solo and all accompanying parts. 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 1 in which vertical lines mark the onsets of each solo note. Score knowledge for musical source separation has also been used in [11] and [12]. Both of these efforts apply a time-varying filter to distinguish the desired audio from its complement. In

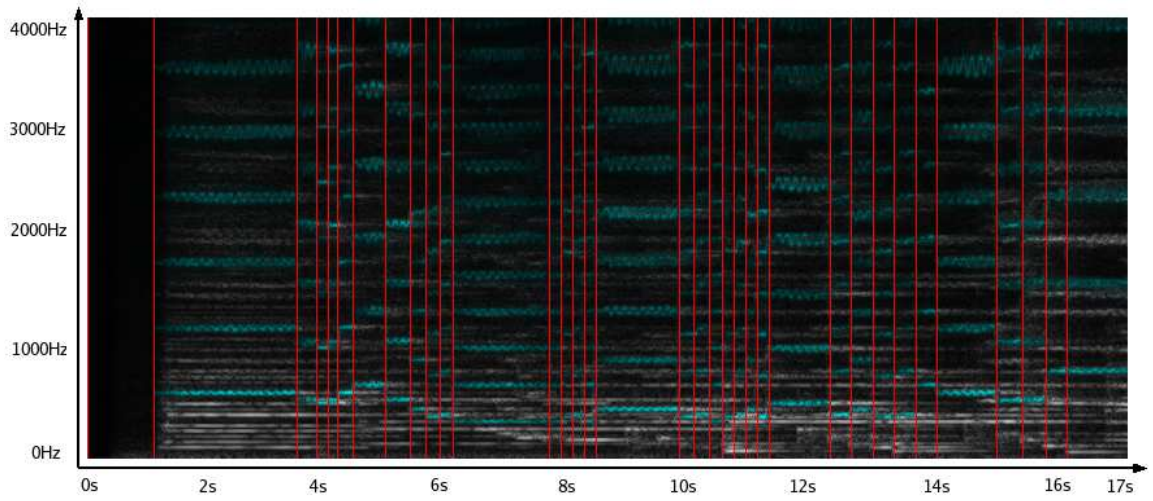

Figure 1: Spectrogram of opening of Samuel Barber Violin Concerto with note onsets for the solo violin marked with vertical lines. A high-resolution version of this same image with the solo part highlighted in blue can be seen at http://xavier.informatics.indiana.edu/~craphael/cmj07.

these efforts, as in ours, the difficulty of identifying the precise time-frequency components one wishes to isolate is the "Achilles' heel" one must inevitably address. Our approach differs from these cited by casting this isolation problem as a one of *classification* and employing appropriate methodology.

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-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 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 well-known idea of *masking* [13], [14]. Using easy-to-create training data synthesized from separate solo and orchestra files, we provide subjective justification for our restricted problem statement. This training data then leads to a principled machine-learning formulation of the problem whose performance we evaluate objectively. We conclude with experiments on data taken from a commercial compact disc in an especially difficult domain — separating the soloist from the orchestra in a concerto setting.

2 STFT Representation

Our approach is based on the short time Fourier transform (STFT) representation of our audio signal. The advantages of this representation are rather obvious for musical signals — much music is composed of notes which are, almost by definition, of limited extent in both time and frequency. Thus, most pairs of notes are supported by entirely disjoint regions of time-frequency space. Even with the STFT, collisions will still occur between harmonics of some notes. However, we believe that other possible signal representations, such as wavelets, share this problem, while the STFT goes as far as any representation can in minimizing the difficulty.

Suppose our audio signal is denoted by

$$x = \ldots, x(-1), x(0), x(1), \ldots$$

We write the short-time Fourier transform of x as $X = X(t,k)$ where

$$X(t,k) = \sum_n x(n) e^{-2\pi i k n/K} w(tH - n) \tag{1}$$

where $k = 0, \ldots, K-1$, $t \in Z$, H is our hop size, and w is the window function which is 0 outside the range $\{-K/2 \ldots, K/2-1\}$. We assume $K = HL$ so that L is the integral number of "hops" needed to traverse the FFT length, K.

Perfect recovery of x from X is exceedingly simple when

$$\sum_t w^2(tH - n) = c \tag{2}$$

for all n and some constant c (see [15] and the references therein for a more detailed discussion). In this case

$$\begin{aligned} x(n) &= \frac{1}{c}\sum_t x(n)w^2(tH-n) \\ &= \frac{1}{cK}\sum_t \sum_{k=0}^{K-1} X(t,k)e^{2\pi ikn/K} w(tH-n) \qquad (3) \\ &= \frac{1}{cK}\sum_t \sum_{k=0}^{K/2} a(t,k)\cos(\phi(t,k) + 2\pi kn/K)w(tH-n) \qquad (4) \end{aligned}$$

where the amplitudes $\{a(t,k)\}$ and phases $\{\phi(t,k)\}$ are taken from $X(t,k)$.

There are several window functions other than the constant window that have the necessary property of Eqn. 2. Among them are the Hanning or "raised cosine" window with $L = K/H = 4$ hops per FFT length, which we use in our experiments.

Eqn. 4 is a rather intuitive description of the original signal as a sum of windowed and translated cosines, whose frequencies are indexed by k and whose translations are indexed by t.

3 Approximate Source Separation

Ideally we wish to decompose our signal x into $x = x_s + x_a$ where x_s corresponds to the solo part and x_a corresponds to the accompaniment. Equivalently, we could seek a decomposition in STFT space: $X = X_s + X_a$, where X_s and X_a are the STFTs of x_s and x_a, though this problem still involves the precise estimation of phase and amplitude for each time-frequency bin of X_s and X_a, subject to the constraint.

Instead we consider the approximations

$$\begin{aligned} X_s &\approx 1_S X \\ X_a &\approx 1_A X \end{aligned}$$

where

$$\begin{aligned} S &= \{(t,k) : |X_s(t,k)| \geq |X_a(t,k)|\} \\ A &= \{(t,k) : |X_s(t,k)| < |X_a(t,k)|\} \end{aligned}$$

and

$$1_C(t,k) = \begin{cases} 1 & \text{if } (t,k) \in C \\ 0 & \text{otherwise} \end{cases}$$

Clearly these approximations are much easier to estimate than the true X_s and X_a since we need only estimate a boolean value for each STFT point, rather than a complex number.

One can appreciate the quality of this approximation by synthetically manufacturing X from known X_s and X_a and listening to the resulting decomposition. To this end, we began with a performance, x_s, of a soloist playing an excerpt from a Mozart violin concerto. We then built the orchestra audio around this performance by first matching both her performance and a prerecorded orchestra performance to a score. We then warped the orchestra recording to synchronize with the solo part using phase-vocoding [16] , [17], [18]. and adjusted the levels to achieve good balance to produce x_a. There are certainly easier ways to produce two synchronized parts, but we already had the machinery set up for the above procedure, and wanted to approximate realistic conditions as well as possible. From the audio files, x_s and x_a, we produced the composite STFT, $X = X_s + X_a$, as well as the two estimates of the separate solo and accompaniment

parts, $\hat{x}_a$ and $\hat{x}_s$, by

$$\begin{aligned} \hat{x}_a &= \text{STFT}^{-1} 1_A X \\ \hat{x}_s &= \text{STFT}^{-1} 1_S X \end{aligned}$$

using Eqn. 3. The three files $x = \text{STFT}^{-1} X, \hat{x}_s, \hat{x}_a$ can be heard at

http://xavier.informatics.indiana.edu/~craphael/cmj07.

The quality of $\hat{x}_s$ and $\hat{x}_a$ was rather surprising to us, sounding, for the most part, quite similar to the original files. This suggests that the effect of our masking operations might not be as significant as one might expect, as has been observed by others in the music processing domain [19].

It is, perhaps, worth noting that our masked versions $1_S X$ and $1_A X$ are not necessarily the STFTs of *any* time signal. This is because the overlapping of windows produces linear constraints the STFT must satisfy. We have no reason to suppose that our masked versions of X would satisfy these constraints. However, the difference between $1_S X$ and $\text{STFT}\hat{x}_s$, is exceedingly small, both in terms of measurable and perceived distance, (similarly for $1_A X$ and $\text{STFT}\hat{x}_a$).

Figure 2 shows the region S as white while the complementary region, A, is colored black. The accompanying web page shows the spectrogram with the two regions colored differently to clearly distinguish them. Perhaps surprising is how much of the STFT is labeled as "solo," including regions seemingly far from any solo harmonics. Part of this chaotic nature of the mask is explained by the spectrogram image. From this image it is clear that the class labels of many of the points are somewhat irrelevant, due to their small contribution to the audio signal.

This experiment demonstrates that the perceptual accuracy of $\hat{x}_a$ and $\hat{x}_s$, thus justifying the use of our approximation. While certainly much easier than trying to estimate X_s and X_a from X, the estimation of our ideal mask is still a difficult problem and will introduce further audio degradation. Thus, the audio

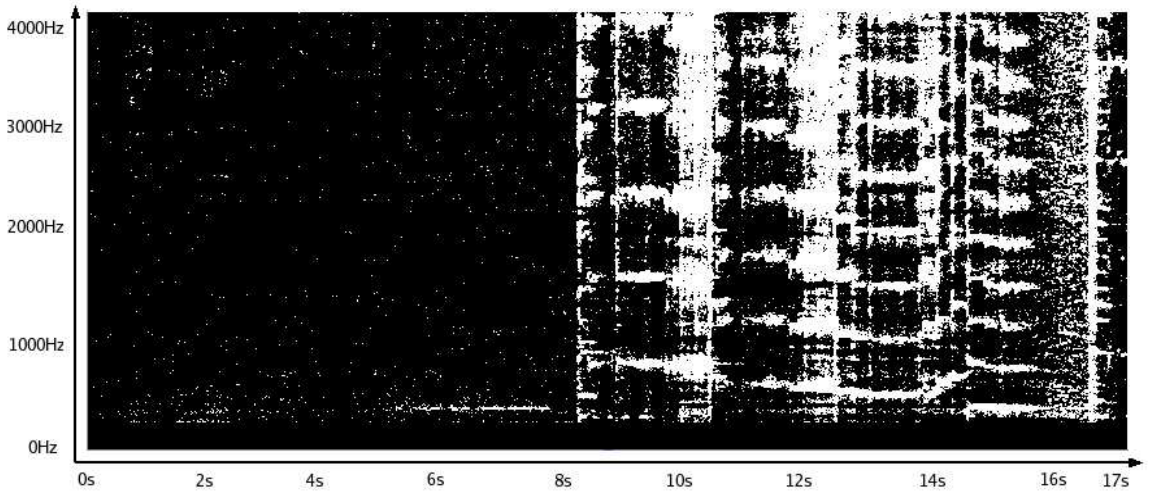

Figure 2: The binary mask indicating, for each point, which part makes the greater contribution. The solo violin is represented as white. See http://xavier.informatics.indiana.edu/~craphael/cmj07 for the spectrogram with color overlayed to express the binary mask.

results should be considered an upper bound on what our masking approach can achieve. The next section develops an approach for estimating the ideal mask.

4 Estimating the Mask

4.1 Classification Trees

Constructing our composite data from unmixed solo and accompaniment parts, as above, leads to principled methods for estimating the ideal mask, as follows. For each point in the composite STFT, $X = X_s + X_a$, we *know* whether X_s or X_a made the bigger contribution. Thus our synthetic spectrum can be viewed as training data for a classifier that attempts to label each point as belonging to S or A. Needless to say, this approach produces voluminous quantities of training data — hundreds of thousands of correctly labeled points for minute-long audio excerpts. With such a large and easily obtainable collection of ground truth, it seems natural to train a classifier to label each STFT point. In addition to the fully automatic construction of the classifier, such an approach allows one to numerically evaluate its success, rather than making subjective judgments of audio quality.

In building our classifier we depart slightly from that presented in the previous section. Many, perhaps most, of the STFT points do not significantly affect our end result, due to their small magnitudes. When building the classifier we eliminate points in $\{|X(t,f)| < T\}$, for some threshold T, since we don't view their labels as meaningful, thus distracting the classifier from its essential task.

With the remaining points we build a tree-structured classifier following the ideas of CART [20]. Our features are derived both from our score match, as well as aspects of the STFT, and consisted of:

vertical dist to closest solo harmonic This feature computes the distance in frequency from the given STFT point to the closest solo harmonic. The feature depends only on the score match. For each STFT point we compute which solo note, if any, is coincident with point and how far the point is, in frequency units from the closest solo harmonic. This feature, by itself, can be used to give somewhat credible results.

vertical dist to closest orchestra harmonic This feature is perfectly analogous to the previous feature, except we consider distances to orchestra harmonics rather than solo harmonics.

distance to closest solo harmonic This feature is also purely a function of the score match. Conceptually we create a binary representation of the solo performance as an idealized spectrogram, containing 1's only where solo harmonic occur. For each STFT point we compute the minimum Euclidean distance to a 1 point over all STFT points of "earlier" times. This feature is useful for detecting points whose energy is mostly due to reverberation of a solo harmonic.

modulus $|X(t,f)|$. High energy points are more likely to be associated with the solo part. We also computed average modulus over local neighborhoods.

rank The percentile ranking of the modulus over a neighborhood of STFT points. The STFT points

associated with the solo tend to be larger in magnitude than their neighbors. This feature was computed over 4x4, 3x3, and 2x2 neighborhoods.

phase coherence One expects that the STFT points composing a harmonic will tend to evolve in time with similar phase advance. This is, in fact, the idea behind the "phase locking" improvements to the phase vocoder [21]. We computed a measure of the degree to which this is true for an STFT point as the empirical variance of the phase advances. We expect this feature will be small on a peak, and especially true for the more closely-recorded solo instrument.

horizontal derivatives Horizontal differences of the STFT moduli were computed in hopes of detecting higher activity for solo harmonics.

We experimented with several other features, but none of these achieved any measurable increase in performance on a validation set.

The classifier is then built according to the usual CART prescription of recursive partitioning, choosing, at each stage, the feature and split point the minimizes the average class label entropy of the two child nodes. We built deep trees, using 680,000 correctly labeled STFT points, splitting tree nodes until a node contains only solo or orchestra points, or until the node has less than 50 points, thus producing thousands of branches. We then prune the tree using traditional CART techniques using an independent validation set of approximately the same size as for training[20].

A portion of the classification results on a separate test set are presented in the accompanying web page, again on the soloist's entrance to the Mozart violin concerto, in which the mistakenly labeled points are indicated with color. The falsely classified points accounted for .025 of the total collection of S and .029 of the total collection of A, out of 680,000 test points. The associated audio reconstructions are not without their merits, but suffer from the discontinuous nature of the purely local processing technique.

4.2 Spatial-Constraint-Based Classification

The decisions of the learned classification tree are mostly based on the distances to solo and orchestral harmonics, as well as local energy in the signal. Clearly these features do not contain enough information to consistently distinguish between solo and orchestra — we doubt any local features can do this. Rather, the separation must be made on less local properties of the signal, which is the approach we move toward in the current section. To this end, we constrained our classifier to estimate masks having a *connected* structure, typical of the masks we seek. This modification identifies two distinct kinds of events occurring within the solo part: note harmonic events and transient events.

For each harmonic of each solo note we consider a rectangular box, B, of sufficient extent to contain the energy generated by that note. The box must extend beyond the "right" edge of the note to include the note's reverberation, and account for our uncertainty in pitch as well. Let t_0, t_e, t_1 denote STFT time indices giving the onset of the note, the onset of the next note, and the latest possible time the note might continue to reverberate. Let the frequency extent be bounded by k_0 and k_1, so $B = \{t_0, \ldots, t_1\} \times \{k_0, \ldots, k_1\}$.

We seek to label all of the points in B as s or a for solo or accompaniment, and write $C(t,k)$ for the label of point (t,k). In the previous section our tree-structured classifier was used to make binary decisions about each point; however, note that our classifier can be used to estimate the *probabilities* of these assignments, e.g. $P(C(t,k) = s|X)$, as the proportion of training examples labeled as s at the terminal node encountered by (t,k). In practice, we smooth these estimates. In this way we use the learned tree as the basis for our data model.

If $I \subseteq B$ is the collection of points labeled as solo, then, assuming independence, the joint labeling, C_B, of all points in B, has probability

$$P(C_B|X) = \prod_{(t,k)\in I} P(C(t,k) = s|X) \prod_{(t,k)\in I^c} P(C(t,k) = a|X) \tag{5}$$

where I^c is the complement if I in B.

To force connectedness of our labeling, we constrain the region I, as follows. For each $t = t_0, \ldots, t_e, \ldots, t_1$ we choose a single (possibly empty) interval, $I_t \subseteq \{k_0, \ldots, k_1\}$, constrained by the requirements

$$I_t \cap I_{t+1} \neq \emptyset \quad \text{when} \quad I_t \neq \emptyset, I_{t+1} \neq \emptyset \tag{6}$$

$$I_{t+1} \subseteq I_t \quad \text{when} \quad t \geq t_e \tag{7}$$

Thus, the sequence of intervals traces out a connected region, $I = \cup_{t=t_0}^{t_1} I_t$, whose vertical extent is non-increasing in the region attributed to reverberation. Subject to the constraints, we seek the set I that maximizes Eqn. 5.

Such a region can easily be identified using dynamic programming. To this end we enumerate the possible intervals, I_t, for each $t \in \{t_0, \ldots, t_1\}$. For each interval I_t we define the data probability

$$D_t(I_t) = \prod_{k \in I_t} P(C(t,k) = s|X) \prod_{k \in I_t^c} P(C(t,k) = a|X)$$

and set $H_{t_0}(I_{t_0}) = D_{t_0}(I_{t_0})$. We then recursively compute the score, $H_t(I_t)$ for $t \in \{t_0 + 1, \ldots, t_1\}$ by

$$H_t(I_t) = \max_{I_{t-1}} H_{t-1}(I_{t-1}) D_t(I_t)$$

where the maximum is over all intervals, I_{t-1} that satisfy Eqns. 6,7. If $I_{t_1}^*$ is the maximizing interval for H_{t_1}, then we can recursively construct the optimal sequence of intervals by

$$I_{t-1}^* = \arg\max_{I_{t-1}} H_{t-1}(I_{t-1}) D_t(I_t)$$

thus producing our optimal sequence of intervals: $I_{t_0}^* \ldots I_{t_1}^*$.

The second type of solo event we identify are *transient* events associated with note onsets. Many instruments produce vertical lines in the spectrogram images at note onset positions, corresponding to widely dispersed spectral energy, before the note settles into its steady-state behavior. While such events are most

obvious in percussive and plucked string instruments, we have observed them in most of the instruments we have studied. These transient events are typically contained within a "thin" and "tall" rectangle in STFT space, $\{t_0, \ldots, t_1\} \times \{k_0, \ldots, k_1\}$, centered in time around the note onset time. Specifically $t_0, \ldots, t_1$ corresponds to around 100 ms. while $k_0, \ldots, k_1$ contains the entire frequency range. We model the transient region as a sequence of *horizontal* intervals $I_k \subseteq \{t_0 \ldots t_1\}$, where $k \in \{k_0 \ldots k_1\}$. The (possibly empty) intervals are constrained by

$$I_k = I_{k+1} \text{ when } I_k \neq \emptyset, I_{k+1} \neq \emptyset$$

thus producing a sequence of rectangles separated by gaps to allow the "free passage" of orchestral harmonics. We seek the collection of rectangles that maximizes

$$H = \sum_{k=k_0}^{k_1} \sum_{t \in I_k} F_v(t,k) - F_h(t,k)$$

where F_v and F_h are two-dimensional filters designed to highlight vertical features (the solo transients) and horizontal features (the orchestral harmonics). Again, this criterion is easily optimized using dynamic programming.

Using an excerpt from a commercial compact disc of Samuel Barber's Violin Concerto, the accompanying web page shows the solo points identified by our spatial-constraint-based classifier colored in purple while the remaining points are colored in blue. The accompanying web page also presents the solo and orchestra audio achieved by inverting the STFT after using the estimated masks. While traces of the unwanted part are occasionally present, we believe these results to be highly promising, especially when considering the challenge of source separation in this orchestral context. It is, of course, not possible to provide any quantitative evaluation of this experiment, since we are not given the "unmixed" solo and orchestra channels.

5 Discussion and Future Directions

Even with our precise score match, 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 accompanying 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 acceptable results.

The most significant contribution of this work is the combination of machine learning techniques with the road map provided by the score match, resulting in a principled way a addressing the desoloing problem. Our technique is generally applicable, in the sense that it does not rely on unrealistic assumptions about the recording process. Beyond that, we have demonstrated a method for training our separating mechanism from real data, as well as numerically evaluating the quality of this separation. Finally, we have offered credible audio results that show the promise score-guided musical source separation.

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. 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 [22]. 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," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[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] Every M. R., (2006) "Separation of Musical Sources and Structure from Single-Channel Polyphonic Recordings," *PhD Thesis, Department of Electronics, University of York*, 2006.

[13] Roweis S., (2000) "One Microphone Source Separation" *Neural Information Processing Systems*, pp. 793–799, 2000.

[14] Bach F. and Jordan M., (2005) "Blind one-microphone speech separation: A spectral learning approach" *Neural Information Processing Systems*, pp. 65–72, 2005.

[15] Zolzer, U. Editor, (2002) "DAFX - Digital Audio Effects," John Wiley and Sons, 2002.

[16] Flanagan, J. L., and Golden, R. M. (1966) "Phase vocoder," *Bell Syst. Tech. J.* , vol. 45, pp. 1493–1509, Nov 1966.

[17] Laroche, J., and Dolson, M., "Phase-vocoder: About this phasiness business," *Proc. IEEE ASSP Workshop on app. of sig. proc. to audio and acous.* New Paltz, NY, 1997.

[18] Puckette, M. (1995) "Phase-locked Vocoder", *IEEE ASSP Conference on Applications of Signal Processing to Audio and Acoustics,* Mohonk, N.Y., 1995.

[19] Li Y., and Wang, D (2006) "Singing Voice Separation from Monaural Recordings," *Proceedings of the 7th International Conference on Music Information Retrieval,* Ed. Roger Dannenberg, Kjell Lemström and Adam Tindale, Victoria, BC, Canada, 176-179, 2006.

[20] Breiman L., Friedman J. H., Olshen R.A. and Stone C. J. (1984) *Classification and Regression Tress*, Monterey, CA: Wadsworth and Brooks/Cole.

[21] Laroche J. and Dolson M., (1999) "Improved phase vocoder timescale modification of audio," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

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