DESOLOING MONAURAL AUDIO USING MIXTURE MODELS

Yushen Han
School of Informaitcs
Indiana Univ.
yushan@indiana.edu

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

ABSTRACT

We describe a new approach to the "desoloing" problem, in which one tries to isolate the accompanying instruments from a monaural recording of a soloist with accompaniment. Our approach is based on explicit knowledge of the audio in the form of a score match – a correspondence between a symbolic score and the music audio, giving the times of all musical events. We employ the familiar idea of *masking* the short time Fourier transform to eliminate the solo part. The ideal mask is estimated by fitting a model to the data, whose note-based components are derived from the score match. The parameters for our probabilistic model are estimated using the EM algorithm.

1 INTRODUCTION

We focus here on the problem of isolating an accompanying instruments from a monaural recording of music for soloist and accompaniment. We call this problem "desoloing." The primary application for desoloing, at least in terms of numbers, would likely be karaoke. Desoloing would produce an accompaniment for any song of interest, thus increasing the range of music on which both singers and listeners could enjoy karaoke.

Our interest in this problem, however, stems from our work with musical accompaniment systems. The idea here seems, at first, painfully close to karaoke, except that the accompanying instruments must follow the soloist, rather than the other way around. This change adds a great deal of complexity to the problem, while also making it attractive to "classical" musicians. Our preferred method of orchestral resynthesis is from actual audio. While commercial orchestral accompaniments are available for some of the solo literature, they tend to be poorly recorded with variable playing. A successful desoloing algorithm would harvest a wide world of beautifully played and expertly recorded orchestras for the accompaniment system. Desoloing serves an MIR need by allowing one to access the "sources" of an audio file independently.

The desoloing problem takes an asymmetric view of the familiar source separation idea, which has received much attention in the signal processing community over the last decade. Much of this work is called "blind" source separation, meaning that one tries to separate the sources with little or no knowledge of their contents [1] [2], [3], [4]. The general area of blind source separation includes several efforts that are explicitly devoted to music audio [5] and [6]. Our framing of the problem is distinct from most work in source separation due to the explicit knowledge we have of the audio — we assume that we are given a symbolic score to the piece of music, giving the pitches and instruments of all notes in the music, as well as a score match, giving a precise correspondence between these notes and the audio file. The present work is enabled by our previous work in orchestral score following with very minor adjustment done manually in case of mismatch. Figure 1 demonstrates this correspondence between score and audio that forms the basis of our approach. A similar problem statement was defined in [7].

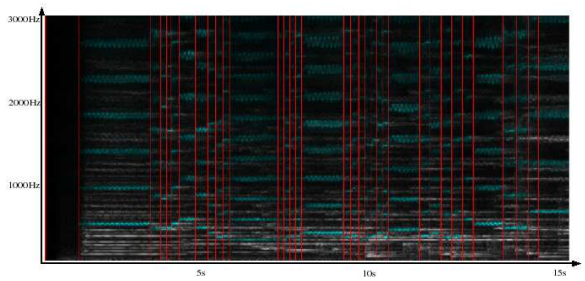

Figure 1. Spectrogram of opening of Samuel Barber Violin Concerto. Vertical lines mark the solo note onset.

While this problem is, no doubt, highly challenging, our particular needs make the goal somewhat more attainable. Any desoloing procedure will almost certainly result in the loss or disfigurement of certain aspects of the orchestral audio we wish to isolate. However, in our accompaniment application, the live soloist will be playing at the precise time-frequency regions where our desoloing procedure does the most damage. Thus, much of the harm done by desoloing will be *masked* by the live soloist. This brings our desoloing into the realm of tractable problems.

2 MASKING IN A TIME-FREQUENCY DOMAIN

Our approach operates on the short time Fourier transform (STFT) X of the audio signal x in the time domain [8]. We use binary masking to decompose X into $X = \hat{X}_s +$

$\hat{X}_a$. $\hat{X}_s$ and $\hat{X}_s$ are our estimates for the solo and the accompaniment. We denote this by

$$\begin{aligned} \hat{X}_s &\approx 1_S X \\ \hat{X}_a &\approx 1_A X \end{aligned}$$

where

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

We define $\hat{A}$ to be the complement of $\hat{S}$.

Convincing audio signals $\hat{x_s}$ and $\hat{x_a}$ in the time domain are reconstructed by $\mathrm{STFT}^{-1}\hat{X}_s$ and $\mathrm{STFT}^{-1}\hat{X}_s$. We include Hann window in STFT with $L = N/H = 4$ hops per FFT length, as it fulfills the constant overlap-add (COLA) constraint for perfect recovery of x from X [13].

It is well-known that perceptually good results can be constructed using the *ideal* mask that can be computed when the sets, S and A are *known*, as when x is artificially constructed by adding together two *known* signals x_s and x_a. For instance, see [9].

Moreover, with *known* x, x_s and x_a, we can derive a percentage that describes how much binary masking we can estimate correctly. This will serve as a quantitative evaluation other than the audible result.

Roweis [14] described the idea of binary masking from a filter bank point of view.

3 MODEL-BASED DECOMPOSITION

3.1 Note-based models

Similar to that of [11], our approach to desoloing relies on a note-based probabilistic model for the magnitude of the STFT, $|X(t,k)|$. Through this model, we decompose the magnitude into two components, one for the soloist and one for the orchestra, via parameter estimation for the model. We then move easily to the classification of each time-frequency point using our decomposition.

Suppose we have a collection of *models*, M, that describes all of the known contributions to our data $|X(t,k)|$. Each model is described by a "template" function q_m, supported on a subset of time-frequency space, D_m, with

$$\sum_{(t,k)\in D_m} q(t,k) = 1$$

The note models will describe the contribution of a given note over a range of frames t, in which the associated q_m would be supported on the frequency bins, k, near the harmonics of the note over the relevant range of frames, t. One could also create models for various other contributions such as the attack and reverberation of a note, etc.

3.2 Statistical assumptions for EM

To employ the expectation-maximization(EM) algorithm, we assume that the magnitude contribution to the spectrogram for each model is given by a collection of independent Poisson random variables $\{Z_m(t,k)\}$ for $(t,k) \in D_m$, as the *hidden* variable in [12], with means $\alpha_m q_m(t,k)$ for some $\alpha_m \geq 0$. Thus, α_m describes the extent to which the contributing event is active, while the average contribution profile, $q_m(t,k)$, is fixed for the model. Furthermore, we assume that

$$|X(t,k)| = \sum_{m\in M} Z_m(t,k) \tag{1}$$

For this model to make sense, the units of $|X(t,k)|$ are scaled so that no significant loss is incurred by regarding the $|X(t,k)|$ as integers, which is consistent with the Poisson assumptions. Strictly speaking, Eqn. 1 cannot be completely correct since, for complex numbers, the sum of the magnitudes is not equal to the magnitude of the sum. However, the assumption is approximately true in the very common case in which one magnitude is much greater than all others. Ellis [10] gives a discussion of this assumption.

3.3 EM algorithm

With the assumptions above, we decompose our spectrum $|X(t,m)|$ by estimating the α_m and q_m parameters using the EM. This algorithm is based on estimating the collection of *hidden* variables, $\{Z_m(t,k)\}$, using the current parameter configuration, and re-estimating the parameters using these estimates. Suppose that α_m^r and q_m^r are the estimates we have after the rth iteration of the algorithm. The E-step of the EM algorithm computes

$$\begin{aligned} C_m^r(t,k) &= E[Z_m(t,k) \mid |X| \,] \\ &= \frac{\alpha_m^r q_m^r(t,k)|X(t,k)|}{\sum_{\mu\in M} \alpha_\mu^r q_\mu^r(t,k)} \end{aligned}$$

$C_m^r(t,k)$ is the estimated contribution to time-frequency point (t,k) given by model m, using our current parameters.

The M-step of the EM algorithm will vary depending on the parameters we are estimating. For the α_m parameters, the M-step would be

$$\alpha_m^{r+1} = \sum_{(t,k)\in D_m} C_m^r(t,k)$$

This is not surprising, since α_m represents our estimate of the total spectral magnitude contribution of model m.

Some of our model templates, q_m, are fixed through the EM iterations. For those that are re-estimated, the M-step will depend on the parametric form of the model.

3.4 Parameters to be estimated

We have experimented with a variety of different models, but, at present, get the best results with a relatively simple configuration. For each note, m in the score, we let $I(m)$ be the range of frames that span the inter-onset interval beginning with the onset of note m and ending with the onset of the following note. This information follows directly from our score match. For each note b, in both

solo and orchestra, we create a model, m, for each frame $t \in I_b$ with domain $D_m = \{(t,0)\ldots,(t,N-1)\}$

$$q_m(t,k) = \sum_h p_h N(k;\mu_h,\sigma_h^2)$$

where $\sum_h p_h = 1$ and

$$N(k;\mu,\sigma^2) = P(k-1/2 < Y < k+1/2)$$

where Y is a normally distributed random variable with mean μ and variance σ^2. For our note models we couple all of the mean values by $\mu_h = h\mu_1(m)$ where $\mu_1(m)$ is the not-necessarily-integral frequency bin for the fundamental frequency of note m. While the peak widths appear to be constant over harmonic number, we achieved better results by allowing the variances σ_h^2 to increase somewhat with frequency. Finally, the p_h constants were taken to be representative of the characteristic frequency profile for the particular instruments.

We have tried to estimate different combinations of these parameters, sharing the parameters in different ways across the entire collection of models. At present, the best assumptions involve estimating only the μ_1 parameter of a solo note for each individual frame in the M-step

$$\mu_1^{r+1}(m) = \frac{\sum_{(t,k)\in D_m} \frac{k}{h_m(k)} C_m^r(t,k)}{\sum_{(t,k)\in D_m} C_m^r(t,k)}$$

where $h_m(k)$ is the harmonic number in $\{1,2,\ldots,\}$ associated with bin k and the pitch of model m. We do not estimate any parameters for the orchestra note models other than the $\{\alpha_m\}$.

The other events we capture are the reverberation of each solo note. The domain of reverberation model m is

$$D_m = \{(t,k) : t \in t_{\text{end}}\ldots, t_{\text{end}}+L_{\text{reverb}}, k \in 0,\ldots,N-1\}$$

where t_{end} denotes the last frame of the solo note, with

$$q_m(t,k) = \sum_h p_h N(k;\mu_h,\sigma_h^2) e^{\lambda(t-t_{\text{end}})}$$

We only estimate the α_m parameter for these models.

4 EXPERIMENTS

Before the work of this paper, the attack or transient phase of a note that distributes spectral energy widely is captured and removed by our *ad hoc* recognizer.

As described in the previous section, a note model, m, can be viewed as a combination of h harmonic components, each of which has its mean and variance. The coefficient p_h is the normalized weight associated with the hth harmonic component of note m. It is not surprising that the configuration of p_h depends on the instrumentation, pitch, and dynamic level of the note. Failure to specify the p_h configuration accordingly leads to a dubious description of the magnitude contribution of the various harmonics. In our experiments, we initialize the configuration of

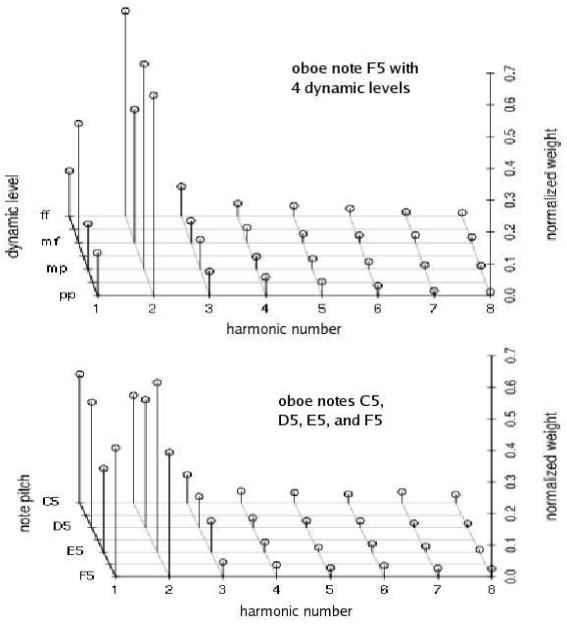

Figure 2. Harmonic weight distribution.

the solo model from an instrument spectrum library using templates trained from a subset of the University of Iowa musical instrument samples. See 4.

In the case that the reverberation of a previous *solo* note contributes, but does not mask the following note, we approximate the decay pattern of the reverberating *solo* note with an exponential over time. Without knowing the acoustic conditions of the recording, we have chosen the parameters experimentally through trial and error. The essence of the "desoloing" problem dictates that we should be generous in our labeling of solo points, since contributions from the solo instrument are readily apparent and undesirable in our results To effect this bias, we employ the following 3 mechanisms:

Initialization The EM algorithm will converge to a local maximum that splits the magnitude of the STFT, $X(t,k)$, into solo and orchestra contributions. We initialize the EM algorithm to expect significantly larger contributions from the solo note s (a ratio of 3:1). Similarly, since the solo reverberation model is longer than the other models in terms of frames, we expect it to consume more spectral energy and assign a larger initial contribution value accordingly.

Harmonic pre-masking Most often when there is a "collision" between a solo harmonic and an orchestra harmonic, most of the spectral data will be due to the solo. In such a case the orchestra model ends up using its free parameters mostly to explain solo energy. To avoid this problem, we identify such collisions using our score and omit the orchestra model for these harmonics.

Masking bias With the final results from the EM algorithm α_m^* and q_m^*. Denote the solo and orchestra

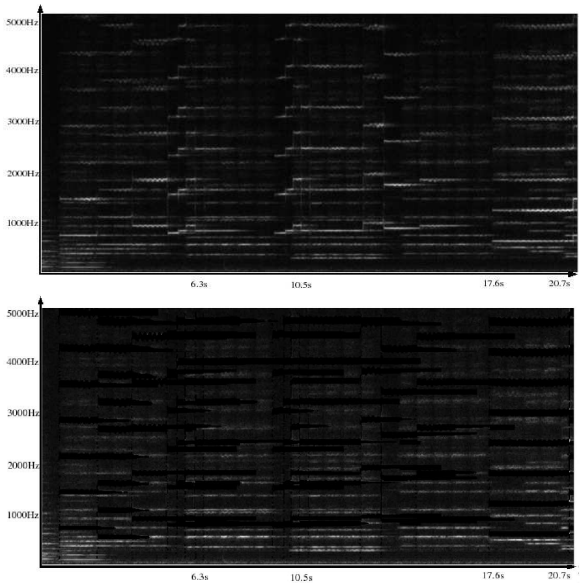

Figure 3. Top: The original spectrogram after transient removal. Bot: spectrogram after desoloing.

models by M_s and M_a. Our solo and orchestra profile estimates are then

$$|\hat{X}_s(t,k)| = \sum_{m \in M_s} \alpha_m^* q_m^*(t,k)$$

$$|\hat{X}_a(t,k)| = \sum_{m \in M_a} \alpha_m^* q_m^*(t,k)$$

We then estimate S by

$$\hat{S} = \{(t,k) : |\hat{X}_s(t,k)| \geq B|\hat{X}_a(t,k)|\} \qquad (2)$$

where B $(0 < B < 1)$ is the constant describing our "generosity" in labling points as solo points, and $\hat{A}$ is the complement of $\hat{S}$.

Our experiments focus on an excerpt from the 2nd movement of Oboe Concerto in C major, K. 314, by Mozart. A desoloed spectrogram is presented in contrast to the original one in Figure 3. The blackened area is corresponding to $S = \{(t,k) : |X_s(t,k)| \geq B|X_a(t,k)|\}$. This audio can be heard at `http://xavier.informatics.indiana.edu/~yushan/desolo_examples.html` in contrast to the original. (A cutoff frequency of 5000Hz is set in order to accelerate the processing.)

A record of a live soloist playing on top of the "desoloed" orchestra is present to demonstrate how the solo will mask the damage done by desoloing.

5 REFERENCES

[1] Bregman, A., *Auditory Scene Analysis.* MIT Press, 1990.

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

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

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

[6] Vincent, E., "Musical Source Separation Using Time-Frequency Source Priors," *IEEE Transactions on Speech and Audio Processing* Volume 14, Issue 1, Jan. 2006 Page(s): 91 - 98

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

[8] Zolzer, U. Editor, *DAFX - Digital Audio Effects.* John Wiley and Sons, 2002.

[9] Li Y., and Wang D., "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.

[10] Ellis D., Chapter 4 of *Computational Auditory Scene Analysis: Principles and Algorithms,* D. Wang and G. Brown eds., Wiley/IEEE Press, pp.115-146, 2006.

[11] Raj B., Smaragdis Ellis D., "Latent Variable Decomposition of Spectrograms for Single Channel Speaker Separation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* pp. 17-20, Oct. 2005.

[12] Bilmes, J.A., "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," International Computer Science Institute, Berkeley, CA, Tech. Rep. TR-97-021, April 1998

[13] Smith, Julius O., *Spectral Audio Signal Processing, March 2007 Version.* Center for Computer Research in Music and Acoustics (CCRMA), Stanford University.

[14] Sam T. Roweis., "One microphone source separation," *Advances in Neural Information Processing Systems* 13. 2001.