INFORMED SOURCE SEPARATION OF ORCHESTRA AND SOLOIST USING MASKING AND UNMASKING

Yushen Han, Christopher Raphael

School of Informatics and Computing, Indiana University Bloomington

yushan@indiana.edu, craphael@indiana.edu

Abstract

A novel technique of unmasking to repair the degradation in sources separated by spectrogram masking is proposed. Our approach is based on explicit knowledge of the musical audio at note level from a score-audio alignment, which we termed Informed Source Separation (ISS). Such knowledge allows the spectrogram energy to be decomposed into note-based models. We assume that a spectrogram mask for the solo is obtained and focus on the problem of repairing the degraded audio. We evaluate the spectrogram as well as the harmonic structure of the music: we either search for unmasked (orchestra) partials of the orchestra to be transposed onto a masked (solo) region or reshape a solo partial with phase and amplitude imputed from unmasked regions. We describe a Kalman smoothing technique to decouple the phase and amplitude of a musical partial that enables the modification to the spectrogram. Audio examples from a piano concerto are available for evaluation.

Index Terms: musical audio source separation BSS score following spectrogram masking Kalman phase estimation

1. Introduction

We address the "desoloing" problem, in which we attempt to isolate the accompanying instruments in a monaural recording of music for soloist and orchestral accompaniment. The motivation is to produce the audio of the accompaniment part for concertos in the "classical" domain as well as the karaoke in popular music, whereas the ultimate goal is to have the orchestra adapt timing to the live player which is a problem we do not discuss there. Nevertheless, the accompanying audio is needed and we offer solutions through our demixing or isolation of the original sources (instruments).

Most past effort in this "source separation" problem treats Blind Source Separation (BSS) problems and assumes little semantic knowledge of the audio content but the independence of the sources [2] [3] or relies on general cues of musical sources rather than the content of the sources [4]. In contrast, we assume explicit knowledge in the form of a score match, which establishes a correspondence between the audio data and a symbolic score representation giving the onset times of all musical events. See Fig. 1 for an example. Such correspondence, known as "score following" or "alignment", initially introduced and developed by Vercoe [1], Dannenberg [16] is the foundation of our approach which we termed *Informed Source Separation* (ISS). Other examples in the category of ISS include Dubnov [8] and Raphael [9]. In our approach, we begin by masking the short time Fourier transform (STFT) in an attempt to "erase" the soloist's contribution. We also based our exploration toward partial-wise phase/amplitude relationship on the work in [10] in which spectrogram magnitude is decomposed into each partial

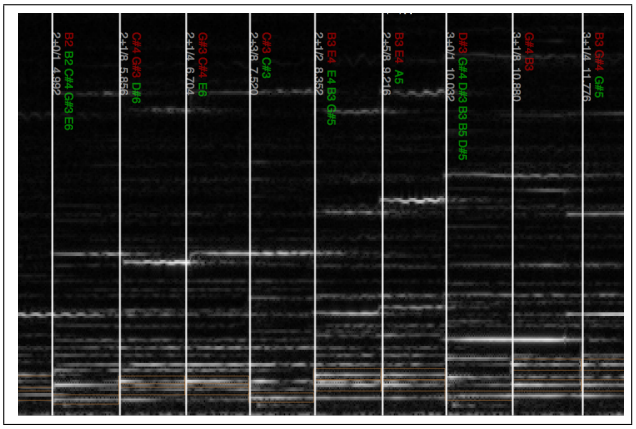

Figure 1: Score following of an excerpt from 2nd mvmt. of Ravel's piano concerto in G (vertical lines mark the piano note onset)

by fitting note-based models. [15] is another effort of spectrogram decomposition in speech. However, our emphasis here is not on estimating the mask or fitting note models, but on employing a novel set of procedures (termed "unmasking") that estimates and then transforms note partials, in which the damage caused by our masking procedure is repaired.

The source separation problem is highly challenging. For the musical accompaniment and karaoke problems one may be able to define success more generously than in other domains - in these cases the live singer or soloist will play at the precise regions in time-frequency space where the most damage has been done through our separation process, thereby hiding some of the inevitable damage.

The structure of this paper is as follows: we briefly formulate the masking problem in sect. 2, followed by our description of partial phase / amplitude estimation in sect. 3 / 4. Such estimating enables our unmasking technique in sect. 5 which is applied in the context of a piano concerto in sect. 6.

2. Spectrogram Masking of the Solo

Given our original audio signal, $x(s)$, we define the short time Fourier transform (STFT) by

$$X(t,k) = \sum_{n=0}^{N-1} x(tH+n)w(n)e^{-2\pi jkn/N}$$

where H is the hop size, N is the window length and w is the window function. We will define our masking operation

in this STFT domain. To do so, we estimate two "complementary" masks, $1_s(t,k)$, and $1_a(t,k)$, taking values in $\{0,1\}$ with $1_s(t,k) + 1_a(t,k) = 1$. These masks are used to isolate the parts of X we attribute to the soloist and accompaniment through

$$X_s(t,k) = 1_s(t,k)X(t,k) \quad (1)$$

$$X_a(t,k) = 1_a(t,k)X(t,k) \quad (2)$$

In other words we label each time-frequency "cell" (t,k) as either solo or accompaniment. Since our focus here is on the *unmasking* problem, we will bias our labeling of each time-frequency cell toward the solo category, since we want to make sure the original soloist is completely removed. Using our score match, it would be relatively easy to simply draw a rectangle around each solo partial while calling the interior of these rectangles our solo mask. Our approach is somewhat more sophisticated, employing special treatment of the wide spectral dispersion associated with note onsets by Ono et al. [19], as well as careful modeling of the steady state partials. However, we will not discuss this mask estimation problem here.

While $X_a(t,k)$ (and $X_s(t,k)$) is, in general, not the STFT of *any* time signal, applying the inverse STFT operation gives perceptually sufficient results with appropriately defined STFT. In particular, if we use a Hann window with $H = N/4$, one can show that applying the STFT inverse to X_a results in the audio signal whose STFT is closest to X_a in the sense of Euclidean distance.

The result of this process elimanates more than the soloist, of course, since the accompanying instruments also contributed to the STFT in the region we have masked out. A possible remedy is the main focus of our paper, treated in what follows.

3. Pastial-wise Amplitude Estimation

In this section we state our technique to decompose spectrogram magnitude into note-based models that can incorporate information from a note sample library. A technique aimed to estimate the solo mask without repairing the damaged orchestra area is documented in [10]. To address those supposedly masked-out partials in collision, we state our adaption of this technique in the following.

3.1. Parameterization of the Music Given the Score

From the score, suppose we have a collection of notes $\mathcal{N}$ in the piece of interest, for a note $n \in \mathcal{N}$, we know its instrumentation $i_n \in \mathcal{I}$ where $\mathcal{I}$ is the set of instruments in this piece and can be further partitioned into disjoined subsets $\mathcal{I}_s$ and $\mathcal{I}_a$ for solo and accompaniment instruments separately.

Moreover, we know the time span of note n: $T_n = \{t_n^{on}, \ldots, t_n^{off}\}$ from the score following. Also, as the note pitch p_n indicates its set of valid harmonics under a certain Nyquist frequency: $\mathcal{H}_n = \{1, \ldots, H_n\}$ [1], we confine the frequency bin span of each partial $h \in \mathcal{H}_n$ to $K_{n,h} = \{k_{n,h}^{low}, \ldots, k_{n,h}^{high}\}$. $K_{n,h}$ implements a band-pass filter to specify a frequency bin span where the contribution from the partial of interest (very likely to be mixed with other partials of close frequencies) is significant in terms of spectrogram magnitude while the spectral energy outside of $K_{n,h}$ is ignored.

Such 2-dimensional, rectangular time-bin support $B_{n,h} = \{(t,k) | t \in T_n, k \in K_{n,h}\}$ specifies a band-passed filter bank over T_n to extract time domain partial $p_h(s)$ from $X(t,k)$ We denote $B_n = B_{n,1} \dot{\cup} \ldots \dot{\cup} B_{n,H_n}$ to be the support for all harmonic components of note n.

We then assume a Normal mixture model for the spectrogram magnitude of an orchestra note n: each harmonic of the note is one Gaussian component in the mixture with normalized weight $\nu_{n,h}$, coupled frequency bin expectation $\mu_{n,h}(t) = h\mu_{n,1}(t)$, and unknown variance $\sigma_{n,h}$. To accommodate the (possibly dramatic) change in amplitude over time of a note, we also introduce a normalized non-negative profile, $\eta_{n,h}(t)$, to outline the frame-wise amplitude of hth partial of nth note.

Strictly, the centroid of each partial may not be precisely coupled by $\mu_{n,h}(t) = h\mu_{n,1}(t)$. But it is approximately true for all the instruments except for piano in our study. To summarize:

- a weight $\nu_{n,h} > 0$ for $\forall(n,h)$ with $\sum_{h \in \mathcal{H}_n} \nu_{n,h} = 1$
- a time support $T_n = \{t_n^{on}, \ldots, t_n^{off}\}$, which is shared among all partials of note n
- an amplitude envelope $\eta_{n,h}(t) > 0$ for $\forall(n,h)$ with $\sum_{h \in \mathcal{H}_n} \eta_{n,h}(t) = 1$
- a frequency bin support $K_{n,h} = \{k_{n,h}^{low}, \ldots, k_{n,h}^{high}\}$
- a frequency bin centroid $\mu_{n,h}(t)$ which reflected the frequency of partial h at t. Among different partials, they are coupled by $\mu_{n,1}(t) = \frac{\mu f_{n,h}(t)}{h}$
- a frequency bin variance $\sigma_{n,h}$ that describes magnitude distribution of partial h over frequency bins with expectation $\mu_{n,h}(t)$ under Normal assumption.

Finally we can define a"template" function $q_{n,h}(t,k)$

$$= \begin{cases} 0, \ \forall(t,k) : t \notin T_n \ or \ \ k \notin K_{n,h} \\ \\ \nu_{n,h}\eta_{n,h}(t) f(k; \mu_{n,h}, \sigma_{n,h}^2); \ \ otherwise \end{cases} \quad (3)$$

where $f(k; \mu_{n,h}, \sigma_{n,h}^2)$ is the probability density function of normal distribution. This parameterization is subjected to normalization to ensure $\sum_{h \in \mathcal{H}_n} \sum_{(t,k) \in B_{n,h}} q_{n,h}(t,k) = 1$ for note n.

3.2. Learning Parameters from Note Samples

To fully use the acoustics knowledge implied in the score (e.g. instrumentation i_n and pitch p_n), we learn the parameters in eq. 3 from note samples of the same instrument and a close, if not the same, pitch. Burred et. al. [17] developed timbre models that capture the instrument characteristics by concatenating notes of the same instrument but different pitches with envelope interpolation. But we favor a pitch-specific model simply because there is no need to compromise f0-dependent information for compactness or generalization in our application as in the instrument identification problem by Kitahara et. al. [18].

The learning process varies according to the number of note samples available. Having a commercial note sample library as well as recorded notes played in different styles and in isolation, which fully covers every pitch that we have in $\mathcal{N}$ in the piece, we sample a probabilistic distribution of $\nu_{n,h}$ for note n of pitch p_n from a collection of note samples of p_n played in different styles. After estimating the frequency of a note sample, we can easily treat the problem of estimating $\sigma_{n,h}$ for each harmonic h as one of a normal distribution with known mean and unknown variance; also, we use an exponential decay function in the envelope $\eta_{n,h}(t)$ for $i_n \in \mathcal{I}_{piano}$ and simply fix $\eta_{n,h}(t) = 1$ for all the orchestra notes. $K_{n,h}$ is chosen as a "confidence interval" according to the estimate of $\sigma_{n,h}$, usually 2-4 frequency bins under $SR = 8000Hz$, 512-point DFT.

[1] H_n does not reach its theoretical maximum due to the smearing effect of changing frequency within a fixed length DFT - some higher partials that cannot be used in our unmasking are skipped

3.3. Statistical Assumption

Our assumption is that the magnitude contribution from each note partial indexed by (n, h) to the spectrogram is raised from a collection of independent Poisson random variables $\{Z_n(t,k)\}$ for $(t,k) \in B_n$ [6]. The expectation of $Z_n(t,k)$ is $\delta_n \sum_h q_{n,h}(t,k)$ where δ_n describes the degree to which $Z_n(t,k)$ contributes to $X(t,k)$.

$$|X(t,k)| = \sum_{n \in \mathcal{N}} Z_n(t,k) \tag{4}$$

Strictly, additivity on the spectrogram only holds for complex entry $X(t,k)$ but we adapt the Max-Approximation discussed in [6] to support the magnitude additivity in t-f cells in eq. 4.

3.4. EM algorithm

With the above assumption, we use EM algorithm to estimate δ_n. At rth iteration

- E-step to estimate $E[Z_n(t,k)||X|]$ using δ_n^r

$$C_n^r(t,k) = \frac{\delta_n^r \sum_{h \in \mathcal{H}_n} q_{n,h}^r(t,k)|X(t,k)|}{\sum_{m \in \mathcal{N}} \delta_m^r \sum_{h \in \mathcal{H}_n} q_{m,h}^r(t,k)} \tag{5}$$

- M-step to re-estimate δ_n^{r+1}

$$\delta_n^{r+1} = \sum_{(t,k) \in B_n} C_n^r(t,k) \tag{6}$$

Intuitively, δ_n is our estimate of the total spectrogram magnitude contribution from note n.

4. Partial-wise Phase Estimation and Transformations

As usually only a subset of partials of a note is damaged by removing the solo partial, we hope to exploited the harmonicity assumption in wind and string instruments supported by Fletcher [13] and Brown [14] to impute the phase of those missing partials in the orchestra. To do so, we first introduce a generic method to decouple the phase and slow-changing amplitude of a band-limited signal in 4.1 which enables our two major tools to "unmask" the damaged spectrogram: harmonic transposition in 4.2 and phase-locked modulation in 4.3.

4.1. Phase Estimation by Kalman Smoothing

In this section we represent our note partial, $p_h(s)$, in terms of a time-varying amplitude and phase:

$$p_h(s) \approx \alpha_h(s) \cos(\theta_h(s))$$

where the time-varying amplitude, $\alpha_h(s)$, is non-negative and varies slowly compared with $p_h(s)$, and the "unwrapped" phase (see Fig. 2) function, $\theta_h(s)$, is monotonically non-decreasing. A more precise review of the slow-changing $\alpha_h(s)$ in a sinusoidal model is given by Rodet [20].

In order to estimate $\alpha_h(s)$ and $\theta_h(s)$ we follow the model of Taylan Cemgil [11] and view the harmonic, $p_h(s)$, as the ouput of a Kalman filter model [21] [22]. To this end we define a sequence of two-dimensional state vectors $\{x(s) = (x_1(s), x_2(s))^t\}$ where $x_1(0)$ and $x_2(0)$ are independent 0-mean random variables with variance γ^2, and the remaining variables follow evolution equation $x(s+1) = Ax(s) + w(s)$ where $\{w(s)\}$ is an independent sequence of 0-mean 2-dimensional vectors with independent components of fixed variance σ^2. A is the rotation matrix, defined in terms of the expected phase advance per sample, ρ, which is directly computable from the nominal frequency of the partial:

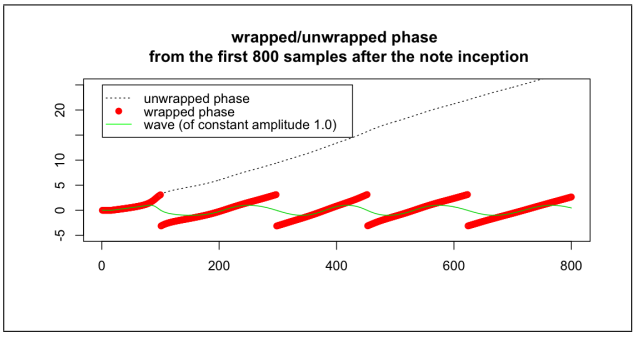

Figure 2: Wrapped and Unwrapped Phase

$$A = \begin{pmatrix} \cos\rho & \sin\rho \\ -\sin\rho & \cos\rho \end{pmatrix}$$

Thus, $x(s)$ is a sequence of vectors that circle around the origin and an approximately known frequency with variable distance from the origin. We then model our observed partial as $p_h(s) = x_1(s) + v(s)$ where $\{v(s)\}$ is another sequence of independent 0-mean variables with variance σ^2.

It is well known that the Kalman filter allows straightforward computation of the conditional distribution, $p(x(s)|\{p_h(s')\})$, and that this distribution is Normal for each value of s. Thus we estimate $x(s)$ by $\hat{x}(s) = E(x(s)|\{p_h(s')\})$. The representation of the partial in terms of amplitude and non-decreasing phase follows from the polar coordinate representation of $\hat{x}(s)$:

$$\begin{aligned} \alpha_h(s) &= \sqrt{\hat{x}_1^2(s) + \hat{x}_2^2(s)} \\ \theta_h(s) &= 2\pi k(s) + \tan^{-1}\left(\frac{\hat{x}_2(s)}{\hat{x}_1(s)}\right) \end{aligned}$$

where each $k(s)$ is chosen to be the non-negative minimal integer value that ensures that $\theta_h(s)$ is non-decreasing.

Note that for phase sequence $\theta_h(s), s \in \{1, \ldots, S\}$, not only the final phase estimate $\hat{\theta}_h(S)$ but also all previous phases estimates are of interest. To get the "best" phase estimation, we need to update the state estimates backward to incorporate the observation that were not "available" at sample s in the forward pass. This motivates Kalman smoothing (see chapter 5 of [22]) which calculates the smoothed phase estimate $\hat{\theta}_h(s)$ recursively backward from the last sample at S.

4.2. Harmonic Transposition

With amplitude $\alpha_h(s)$ and phase $\theta_h(s)$ decoupled from hth harmonic of a note, we are ready to "project" one harmonic into a different harmonic while maintaining the harmonicity between the source and the destination. Supposing we estimated the unwrapped phase of the ith harmonic as $\theta_i(s)$, the "projected" phase sequence at jth harmonic is given by $\tilde{\theta}_j(s) = \frac{j\theta_i(s)}{i}$ and the resulting jth harmonic by

$$\tilde{p}_j(s) = \tilde{\alpha}_j(s) \cos\left(\frac{j\theta_i(s)}{i}\right) \tag{7}$$

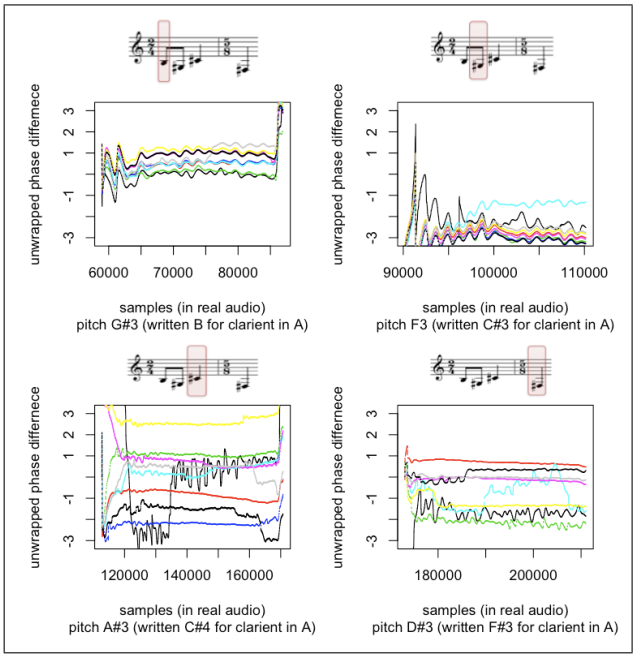

Figure 3: Unwrapped Phase Difference

where $\tilde{\alpha}_j(s)$ is either known or imputed amplitude at jth harmonic. In this work, we usually have an estimate of $\tilde{\alpha}_j(s)$ from the decomposition of spectrogram magnitude described in 3.

Our harmonic transposition exploit such "harmonicity" between partials, which is a well-studied phenomenon. Early work mainly by Fletcher showed that frequencies of the partials in"the middle portion of the tone" of string instrument are integral multiples of the fundamental frequency by using sonograph and also derived that partials of string and wind instrument are "rigorously locked into harmonic relationship" [13]. By using single frame approximation on a variety of digital samples, Brown concluded that "continuously driven instruments such as the bowed strings, winds, and voice have phase-locked frequency components with frequencies in the ratio of integers to within the currently achievable measurement accuracy of about 0.2%" [14].

To demonstrate such harmonicity in our framework, we focus on the "projection" of the unwrapped phase $\theta_i(s)$ from partial i to partial j by

$$\theta_{i,j}(s) = \frac{j\theta_{h_1}(s)}{i} \tag{8}$$

By "projecting" the phase of different partials to a common harmonic, we can examine such phase relation on a variety of orchestra instruments. We can visualize *pairwise phase difference* $\theta_{i,1}(s) - \theta_{j,1}(s)$ at the fundamental for any $i \neq j$. Fig. 3 shows the *pairwise phase difference* for the first 4 notes from a performance of the first movement of Stravinsky's Three Pieces for Clarinet Solo. The salient message from this plot is: the *pairwise phase difference* is in a very small range (mostly $(-\frac{\pi}{2}, \frac{\pi}{2})$) and never drifts away over the entire note; the error (including measurement error and true difference) is not accumulative. This supports our approximation of phase coherence.

Piano and other impulsively driven instruments such as strings played pizzicato are counter-examples whose partials deviate from integer ratios due to the stiffness of the string [14].

4.3. Phase-locked Modulation

In addition to the partial-wise relationship, we want to exploit timewise similarity in terms of phase and amplitude within one note.

Suppose we have a partition $T_1 = \{s_1, \ldots, s_k - 1\}, T_2 = \{s_k, \ldots, s_2\}$ on the sample indices $T = \{s_1, \ldots, s_2\}$ of the sustaining part of a reasonably long orchestra note, we can only observe the unwrapped phase sequence at $\theta_h(T_1)$ but $\theta_h(T_2)$ is missing. We can impute $\theta_h(T_2)$ sequentially by

$$\theta_h(s_k+n) = \theta_h(s_k+n-1)+\theta_h(s_1+1+n)-\theta_h(s_1+n) \tag{9}$$

for any $0 \leq n \leq s_2 - s_k$. We omit the formula to obtain $\theta_h(T_1)$ if we observe $\theta_h(T_2)$.

This operation reserves the phase advance per sample in T_1 and applies such $\Delta\theta_h(T_1)$ cyclically to T_2. This is similar to the phase vocoder except for that we are doing it on the sample level rather than frame level. For a long enough time span T_1, we are capturing the pattern of frequency fluctuation in $\theta_h(T_1)$. To synthesize a segment of a partial, we also need the amplitude envelope over T_2. A simple solution is to reuse the average amplitude α_h over T_1 (with some minor modulation) to "sustain" a note through the end of T_2. If the orchestra note is holding for quite long, which is common in some orchestration, we are effectively synthesizing the sustaining part of the partial.

5. Spectrogram Unmasking

In an attempt to fix the damage caused by desolo, we examine the spectrogram with a focus on areas where the accompaniment notes (harmonics) are damaged.

Our assumption is that there is information redundancy in terms of phase and amplitude between the "observable" partials (i.e. not significantly overlapped by the solo or an accompaniment instrument of a different family) and damaged partials. Our hope is to "copy and paste" musical partials from the observable area to the damaged area with some necessary transformations that exploit those redundancy to maintain the consistency between the observable and the damaged. These procedures can be automated by analyzing the texture of the music from the score and testing the soundness of remaining partials on the desoloed spectrogram. We call this process *unmasking* in which the masked-out solo regions will be recovered.

In the type of music that we (and many solo musicians) are mainly interested in, for instance, a piano concerto, it is common that a string section may double the solo instrument at the unison, fifth, or octave in either direction. In these cases, masking out the solo part usually results in many damaged partials in the orchestra since consonant intervals mean more partials are likely to share the same frequencies. With this in mind, we use some heuristics to create an algorithm to automatically perform the two partial-wise transformations developed in 4.2 and 4.3. Since the texture of the music can be highly complex, we reconstruct a somewhat "generic" scenario for illustration of this algorithm in Fig. 4. The 1-bar score in the figure is a reduction from a piano concerto where the piano part is frequently doubled by the lower string sections.

Supposing we have obtained solo mask $1_s(t,k)$, a damaged region $B^d_{n,h} \subseteq B_{n,h}$, a template $g_{n,h}(t,k)$ and an amplitude estimate δ_n from section 2 and 3 for a damaged partial h of note n, we summarize our *heuristic* algorithm below:

First, we need to evaluate the damage. If $\sum_{(t,k)\in B^d_{n,h}} g_{n,h}(t,k) \ll \sum_{(t,k)\in B_{n,h}} g_{n,h}(t,k)$, we leave it as intact; otherwise we need to repair it. Specially, if

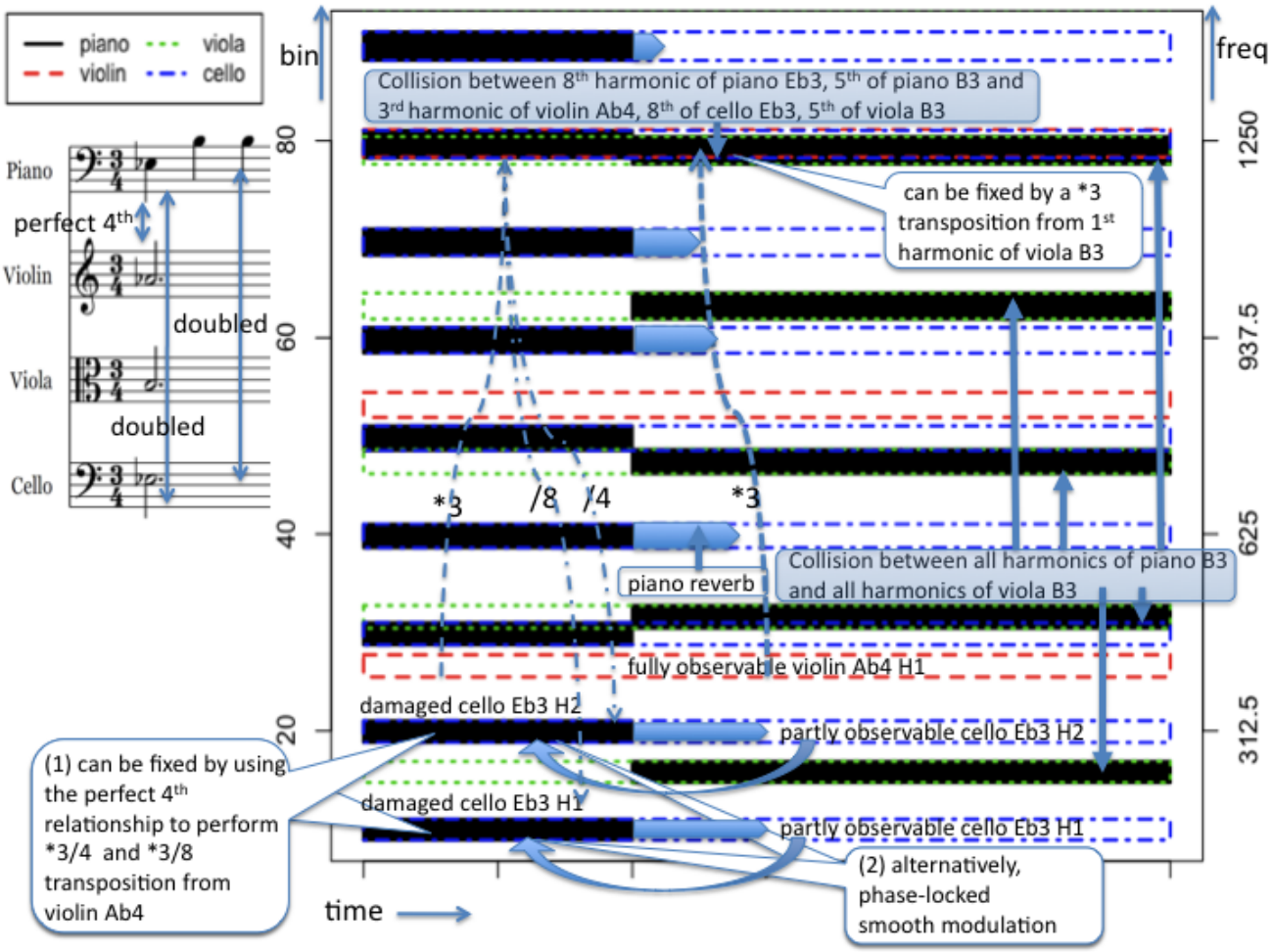

Figure 4: Evaluating Desolo Damage and Possible Fix Using Both Score and Spectrogram

undamaged part $B_{n,h} \setminus B^d_{n,h}$ is a narrow band-limited "strip" (e.g. a single frequency bin), we need to "expand" the solo mask to remove those initially deemed "undamaged" f-t cells as well because such residue tends to create artifact "musical noise" whose suppression deserves treatment, mostly from speech enhancement [12]. After such extra "masking", we use $B^u_{n,h} \subseteq B_{n,h}$ to denote the remaining undamaged region.

Second, since $B_{n_1,h_1} \cap B_{n_2,h_2} \neq \emptyset, n_1 \neq n_2$ for possibly many different note partials contributing energy to the same region, we choose one damaged orchestra partial (n,h) to repair: $\underset{(n,h)}{\operatorname{argmax}} \sum_{(t,k)\in B_{n,h}} \delta_n g_{n,h}(t,k)$ assuming Max-Approximation that only one signal dominates in each time-frequency cell [6].

Third, in the score we look for consonant intervals such as octaves, perfect 5th and perfect 4th in the hope to find an observable partial whose frequency is in a relatively simple ratio to the damaged one waiting to be "transposed" to. We call this partial, if exists, a *candidate*. Usually more than one candidate exist. Large modulus value, simple frequency ratio and identical instrumentation are factors that we favor in choosing the best candidate without creating artifacts. Thus, harmonic transposition can be performed vertically on the spectrogram (e.g. from 3rd to 5th harmonic of viola note B3 in Fig. 4) if the duration of the candidate partial covers that of the damaged area.

Forth, when there is no candidate partial for the partial indexed by (n,h), if there exists a partial (m,i) whose time support of its undamaged portion $T^u_{m,i}$ is adjacent to the damaged duration $T^d_{n,h}$ and whose frequency bin support $K_{m,i}$ satisfies $K^d_{n,h} \subseteq K_{m,i}$ we can perform phase-locked modulation with differenced phase sequence estimated from $B^u_{m,i}$ to $B^d_{n,h}$. The 2 cello partials in Fig. 4 are repaired this way.

Occasionally, we are unable to perform either transformation and label the damaged partial as such.

6. Experiment Results

We experiment with an excerpt of 45 seconds from the 2nd movement of Ravel's piano concerto in G major.

Table 1 lists a breakdown of the number of partials and the number of harmonic transpositions and phase-locked modulation that our algorithm performed. The second column gives the number of partials that have significant spectral energy below Nyquist frequency at SR=8000Hz. Among them, many need to be repaired depending how frequently they collide with the piano. The last column, "unable to fix" gives the number of occurrences that no undamaged orchestra partial is available to estimate phase from. We relax on that the 4 sections of string instruments can be used to repair each other by harmonic trans-

	note	partial	tran. from	tran. to	modulation	unable to repair
oboe	20	85	1	1	0	1
clarinet	6	18	3	3	0	0
flute	6	18	0	0	0	0
violin1	5	42	14	9	0	0
violin2	11	107	34	24	24	2
viola	16	160	33	41	64	5
cello	12	120	43	50	22	6

Table 1: Instrument breakdown of partials being repaired

position but do not allow any harmonic transposition between two different instruments in the woodwind family. This is because the oboe is sharper than the other two in this excerpt. At the end the most of damaged partials are fixed in some way. We also notice that the woodwinds are less damaged because the notes are very high pitched and too loud to yield to the solo piano at their time-frequency region, while the lower string instruments are frequently damaged.

The original, desoloed-but-unrepaired and repaired audio are available at our demo website [24] to evaluate the solo mask and improvement from unmasking. Plots in color giving a breakdown of the partials on the spectrogram are also available.

7. Conclusion, Evaluation and Future Work

Instead of merely extracting one source (instrument) of sound from the mixture, we distinguish our proposed ISS method from other known source separation methods by our explicit *repair* stage that addresses the audio degradation caused by the separation procedure. This stage significantly enhances the perceptual audio quality and boosts performance measurement such as distortion due to interferences proposed by Vincent et al. in [5]. That the reconstructed note sounds plausible for some orchestra instruments suggests that the partial-wise phase/amplitude relationship is a potentially fruitful topic to investigate.

At this stage, we admit that the comparison of our method of "unmasking" with other missing data inference techniques such as [23] is not available and hence is our future work. An ideal evaluation of any method of solo/orchestra separation requires a "ground truth" of the two sources recorded separately and an artificial mix of the two. However, such "ground truth" is almost away absent in the real case and the evaluation is mainly subjective. To explore this open-ended problem, we use interactive visualization and auralization to experiment with the "reconstructed" partials under different settings. Using a music sample library, we can artificially construct ground truth according to the score while maintaining the texture of the music of interests. Some early exploration can be found at [24] as well.

8. References

[1] B.L. Vercoe: "The Synthetic Performer in the Context of Live Performance," *in Proc., International Computer Music Conference*, 1984, Paris, pp. 199-200.

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

[3] Belouchrani, A. Abed-Meraim, K. Cardoso, J.-F. Moulines, E.: "A Blind Source Separation Technique Using Second-Order Statistics," *IEEE Trans. on Signal Processing*,1997, Vol 45; No. 2, pages 434-444

[4] E. Vincent: "Musical Source Separation Using Time-Frequency Source Priors," *IEEE Trans. on Speech and Audio Processing* , Vol. 14, Iss. 1, Jan. 2006 pp. 91 - 98.

[5] E Vincent, R Gribonval, and C Fevotte: "Performance measurement in blind audio source separation.", *Audio, Speech, and Language Processing, IEEE Transactions on* 14(4):14621469, 2006.

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

[7] A. S. Bregman: *Auditory scene analysis*. MIT Press: Cambridge, MA, 1990.

[8] S. Dubnov:"Optimal filtering of an instrument sound in a mixed recording using harmonic model and score alignment," *Proc. of Intl. Computer Music Conf.* 2004, Miami.

[9] C. Raphael: "A classifier-based approach to score-guided source separation of musical audio." *Computer Music J.* vol. 32, no.1 (Mar. 2008), pp. 51-59. 2008.

[10] "removed for reviewing"

[11] A. T. Cemgil, S. J. Godsill: "Probabilistic Phase Vocoder and its application to Interpolation of Missing Values in Audio Signals." Antalya/Turkey, 2005. EURASIP.

[12] S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoust., Speech, Signal Process.*, vol.27, pp. 113-120, Apr. 1979

[13] H. Fletcher: "Mode locking in nonlinearly excited inharmonic musical oscillators," 64, pp. 1566-1569 *J. Acoust. Soc. Am.*, 1978.

[14] Judith C. Brown: "Frequency ratios of spectral components of musical sounds," *J. Acoust. Soc. Am*. 99, 1210 (1996).

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

[16] R. Dannenberg and C.Raphael: "Music Score Alignment and Computer Accompaniment," *Communications of the ACM, 49(8) (August 2006), pp. 38-43.*

[17] J. Burred, A. Rbel and X. Rodet: "An Accurate Timbre Model for Musical Instruments and its Application to Classification," *Proc. Intl. Workshop on Learning the Semantics of Audio Signals (LSAS)*, Athens, Greece, Dec. 2006.

[18] T. Kitahara, M. Goto, and H. G. Okuno: "Musical instrument identification based on F0-dependent multivariate normal distribution," *ICME '03: Proc. of the 2003 Intl. Conference on Multimedia and Expo - Vol. 3* pp. 409-412, 2003.

[19] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama: "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," *Proc. EUSIPCO.*, August 2008

[20] Xavier Rodet: "Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models," *IEEE Time-Frequency and Time-Scale Workshop 97*, Coventry, Grande Bretagne, aot 1997

[21] R. E. Kalman: "A New Approach to Linear Filtering and Prediction Problems," *Transaction of the ASME - Journal of Basic Engineering*, 35-45. March 1960.

[22] R.L. Eubank: *A Kalman Filter Primer*, Chapman & Hall/CRC, 2005.

[23] J Bouvrie and T Ezzat.: "An incremental algorithm for signal reconstruction from short-time fourier transform magnitude." *9th Intl. Conf. on Spoken Language Processing,* 2006.

[24] `http://129.79.223.1/SAPA2010`