

DESOLING MONAURAL AUDIO USING MIXTURE MODELS

For ISMIR 2007 @ Vienna, AT

By Yushen Han, Christopher Raphael

Motivation

A lack of a live accompanist is a common problem that musicians, music students and amateurs suffer from. Our musical accompaniment system is its solution in which the accompanying instruments must follow the soloist, rather than the other way around. In this system, a recording of the music performed by full orchestra or chamber ensemble or jazz band, to which the user would contribute the missing solo part, is the essential material for resynthesizing the accompaniment part.

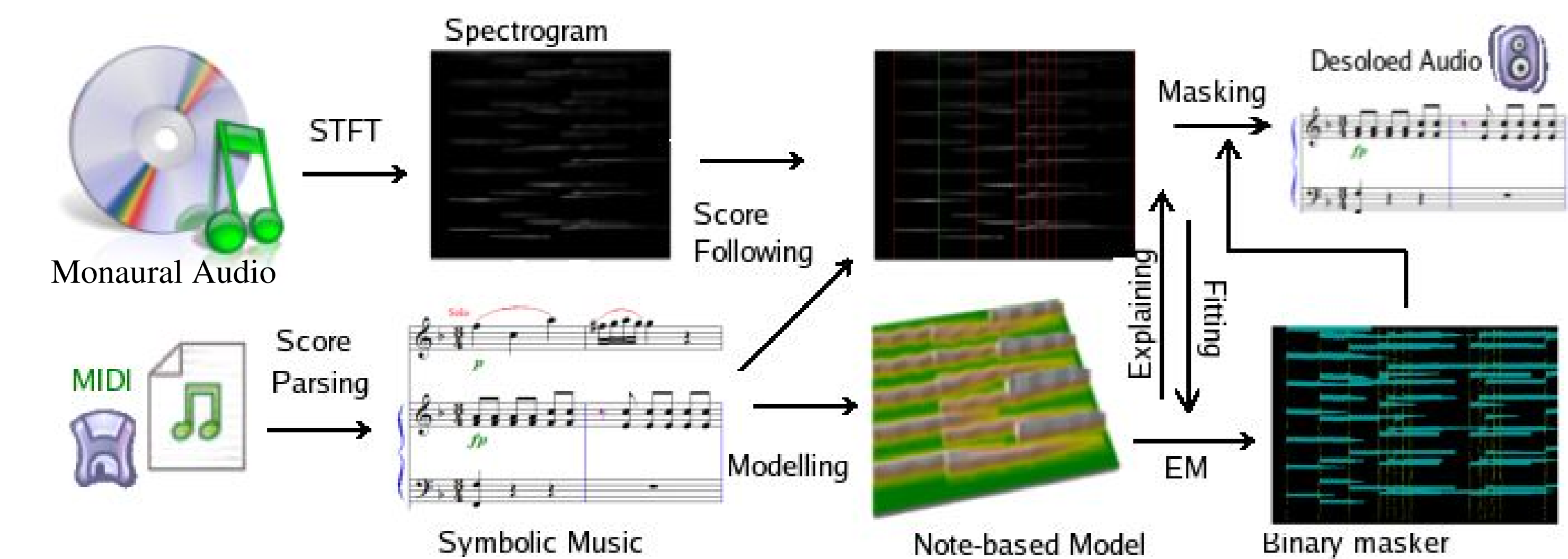
While commercial orchestral accompaniments are available for some of the solo literature, they tend to be poorly recorded with variable playing. We focus here on the problem of isolating a recording of the accompanying instruments from a complete monaural recording of music for soloist and accompaniment. We call this problem "desololing" and aims to harvest a wide world of beautifully played and expertly recorded orchestras for the accompaniment system.



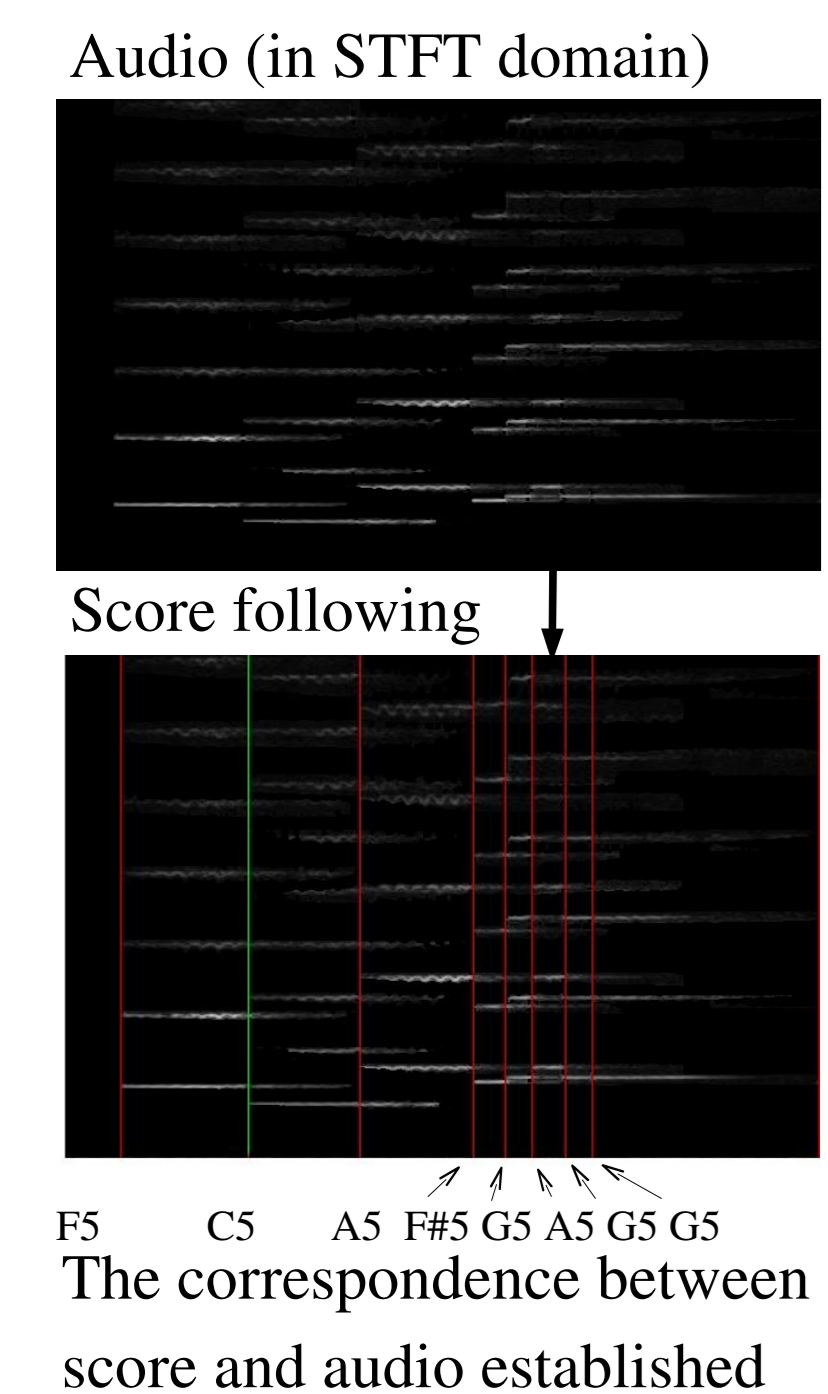
The well-known Music Minus One project provides the printed music with recordings of the piece's orchestral or piano accompaniments minus the soloist.

Approach

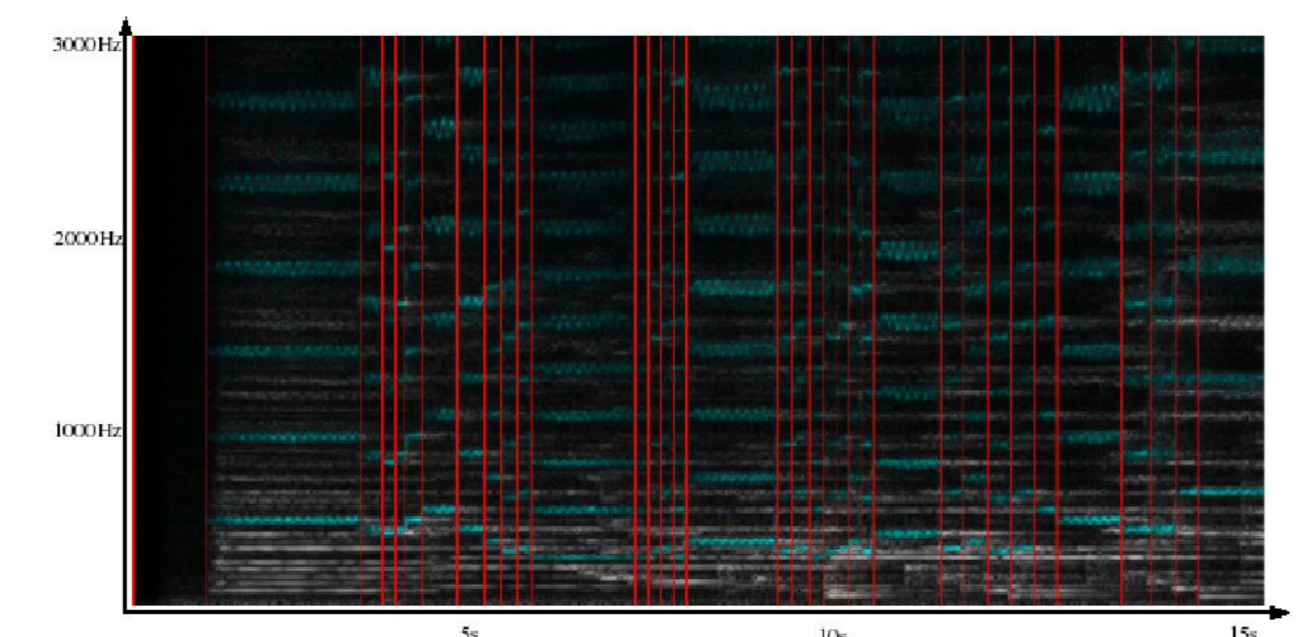
Our approach is based on explicit knowledge of the audio in the form of a score match -- a correspondence between a symbolic score and the music audio, giving the times of all musical events. We employ the familiar idea of masking the short time Fourier transform to eliminate the solo part. The ideal 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.



Previous Work - Score Following



The present work is enabled by our previous work in orchestral score following, with very minor adjustment done manually in case of mismatch. We could make a good use of the information of the score in order to initialize note models in the spectrogram later. We are mostly interested in the onset time and the pitch of each note.



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.

Masking in the Time-Frequency Domain

We use binary masking to decompose the audio signal X in STFT domain into

$$X = X_s + X_a$$

X_s and X_a are the audio for the solo and the accompaniment in STFT domain.

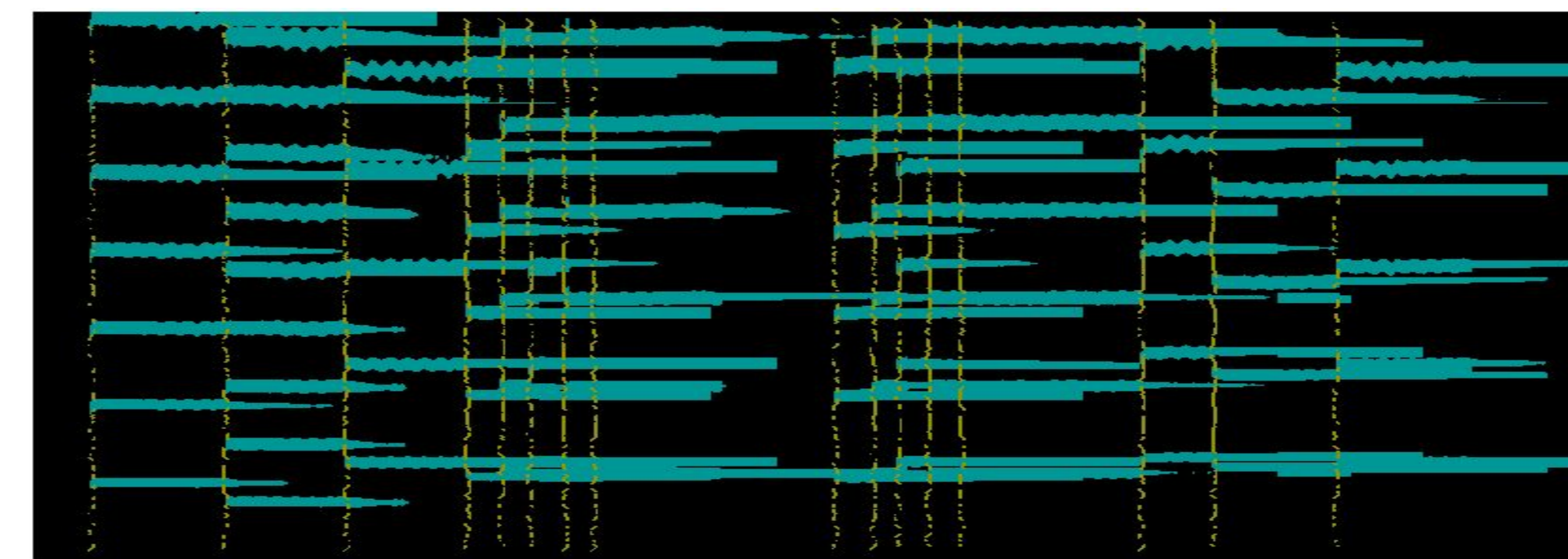
Then we employ the following masker in the 2-dimensional time-frequency domain

$$\hat{X}_s \approx 1_S X$$

$$\hat{X}_a \approx 1_A X$$

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

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

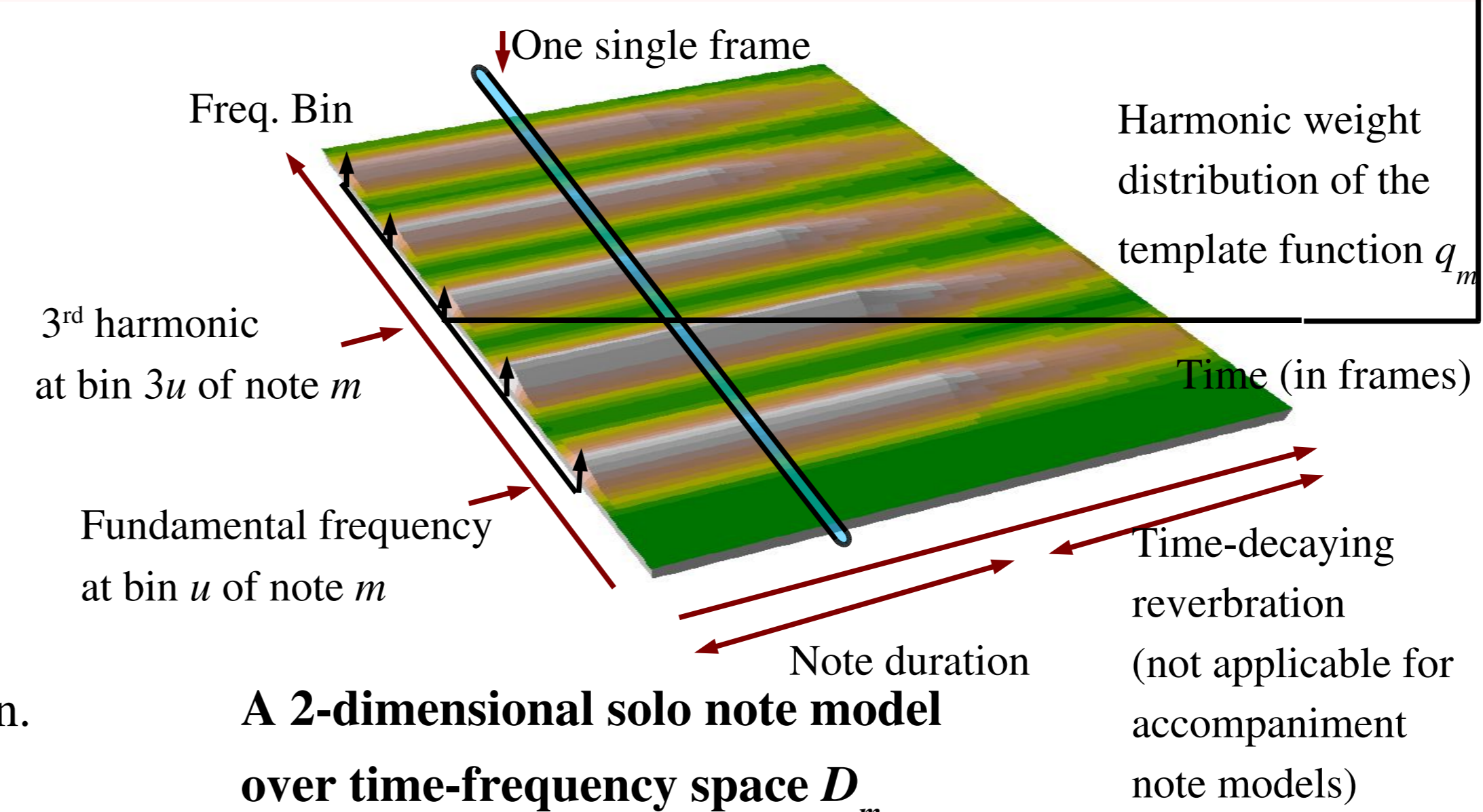


An estimated masker. The blue area masks the solo notes, the yellow area removes the attack and transient behavior of the solo notes (by our *ad hoc* recognizer), while the black area indicates what we have left (the accompaniment part) after the masking.

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 in time domain is artificially constructed by adding together two known signals x_s and x_a .

Model-based Decomposition

Our approach to desololing 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 $\sum_{(t, k) \in D_m} q_m(t, k) = 1$. 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 .

Statistical Assumption:

$Z_m(t, k)$ for (t, k) in D_m the magnitude contribution to the spectrogram for each model with means $a_m * q_m(t, k)$ ($a_m \geq 0$ describes the extent to which the contributing event is active)

Expectation-maximization (EM) algorithm:

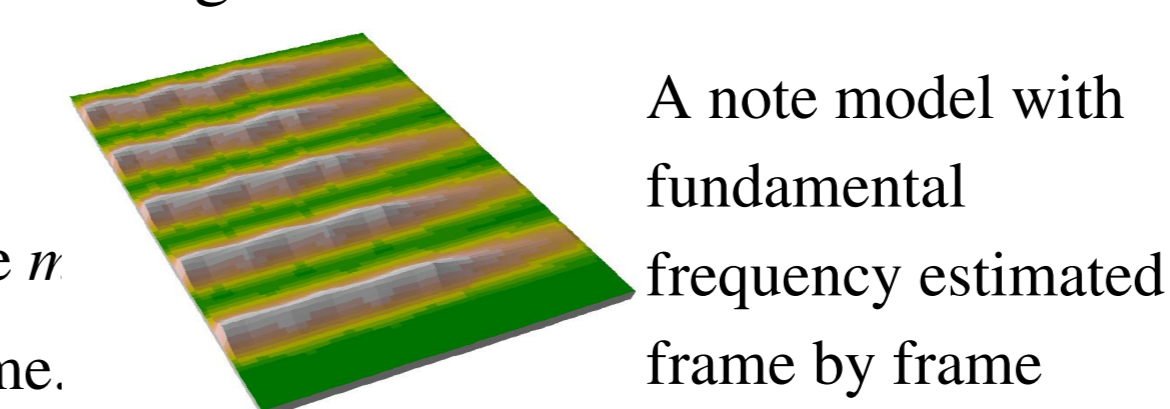
We decompose our spectrum $|X(t, k)|$ by estimating the a_m and q_m parameters using the EM.

Suppose $a_m^r * q_m^r$ are the estimates we have after the r th iteration of the algorithm.

The **E-step** $C_m^r(t, k) = \frac{E[Z_m(t, k) | |X|]}{\sum_{\mu \in M} \alpha_\mu^r q_\mu^r(t, k)}$ estimates the contribution to time-frequency point (t, k) given by model m , using our current parameters.

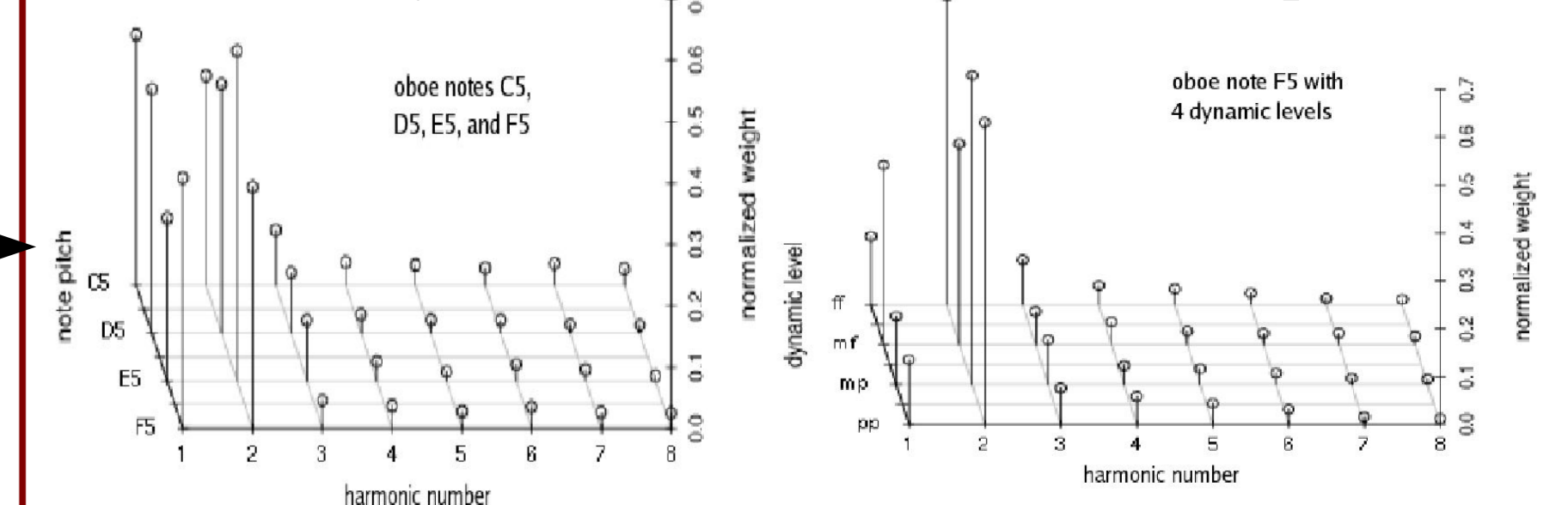
The **M-step** $\alpha_m^{r+1} = \sum_{(t, k) \in D_m} C_m^r(t, k)$ estimates a_m , the total spectral magnitude contribution of model m .

Since we couple the harmonics by $u_h = h * u$, where h is the harmonic number of note n we also estimate the fundamental frequency u of a solo note for each individual frame.



Experiments

Since each note model m , is a combination of h harmonic components, each of which has its mean and variance, we associate the h th harmonic component of note m with a normalized weight p_h . The configuration of p_h depends on the instrumentation, pitch, and dynamic level of the note. In our experiments, we initialize the configuration of the solo model from an instrument spectrum library using templates trained from a subset of the University of Iowa musical instrument samples.

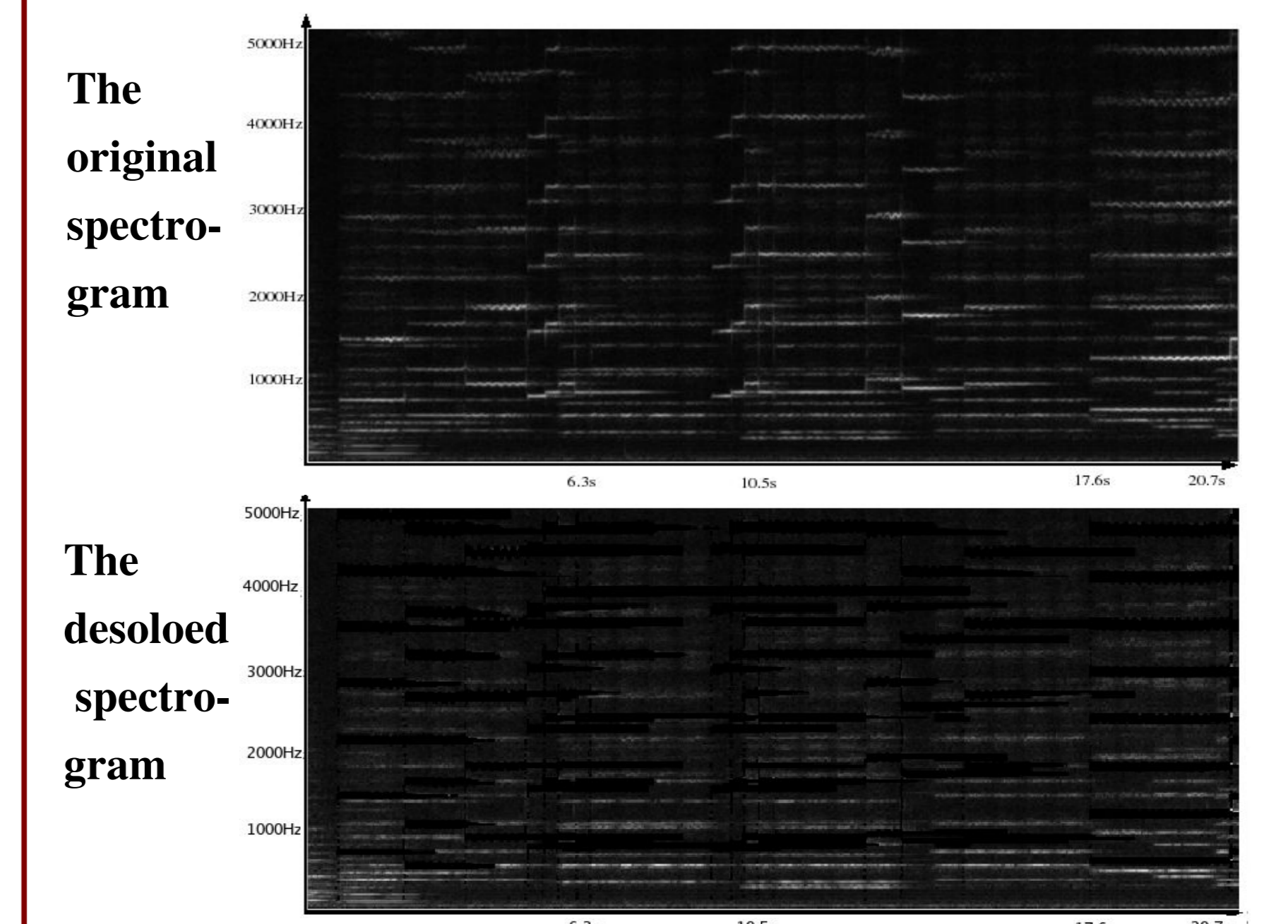


Since contributions from the solo instrument are readily apparent and undesirable in our results, we should be generous in our labeling of solo points. To effect this bias, we employ the following 3 mechanisms:

- Initializing the EM algorithm to expect significantly larger contributions from the solo note
- Omitting the orchestra model where there is a "collision" between a solo harmonic and an orchestra harmonic
- Marking solo points with a favorable bias in the final estimate from the EM algorithm

Results

On an excerpt from 2nd movement, Mozart Oboe Concerto KV314



(A cutoff frequency of 5000Hz is set to accelerate the process.)

With the phase vocoding technique, the "desolored" excerpt is warped to match the oboe audio by a soloist. In the mixed audio, the defiguration in the orchestra audio caused by our isolation procedure had been largely masked by the live soloist.