

# MULTIPITCH ESTIMATION USING SPARSE IMPULSE DISTRIBUTIONS AND INSTRUMENT SPECIFIC PRIORS

Gautham J. Mysore

Center for Computer Research in Music and Acoustics  
Stanford University

Paris Smaragdis

Advanced Technology Labs  
Adobe Systems Inc.

## 1. INTRODUCTION

We present an algorithm for the concurrent estimation of the relative pitch tracks of multiple instruments in a sound mixture. A relative pitch track and a timbral signature is concurrently estimated for each instrument in the mixture. The estimation is carried out by performing multiple simultaneous deconvolutions using shift-invariant probabilistic latent component analysis on a time-frequency magnitude representation (constant-Q transform) of the sound mixture. The deconvolutions yield a kernel distribution (spectral signature) and impulse distribution (from which the relative pitch track is extracted) for each instrument in the mixture. Entropic prior distributions are used on the impulse distributions in order to make them sparse. Spectral characteristics of certain instruments are specified using instrument specific prior distributions on the kernel distributions. Additionally, a Kalman filter type smoothing is used to enforce temporal continuity of pitch tracks.

## 2. CONSTANT-Q TRANSFORM

The constant-Q transform uses a logarithmic spacing of the frequency axis. This implies that the spacing between any two harmonics of a harmonic series is constant with respect to the frequency axis, regardless of the fundamental frequency. Due to this, the spectrum of a given note of music can be shifted up or down in frequency to yield the spectrum of another note of music. This is in contrast to a standard magnitude spectrogram which uses a linear spacing of the frequency axis. We can therefore view the constant-Q transform of a sequence of notes as an approximately constant spectral pattern with different shifts along the frequency axis at different instants of time.

## 3. SHIFT-INVARIANT PROBABILISTIC LATENT COMPONENT ANALYSIS

The constant-Q transform of a sequence of notes is modeled as a probability distribution,  $P(f, t)$ . We model the constant spectral pattern as a probability distribution called the kernel

distribution,  $P_K(\tau_f)$ . This gives us a spectral signature for a given instrument. We model the shifts in frequency by another probability distribution called the impulse distribution  $P_I(f', t)$ . The constant-Q transform of the given sequence of notes is therefore modeled as the convolution of the kernel distribution and the impulse distribution.

We model a sound mixture that contains a number of different instruments as the sum of their constant-Q transforms, since time-frequency magnitude representations are approximately additive. Each of these constituent constant-Q transforms are in turn modeled as the convolution of a kernel distribution  $P_K(\tau_f|z)$  and an impulse distribution  $P_I(f', t|z)$ , where  $z$  is a latent variable that represents the  $z$ -th instrument. We model the relative proportions of each instrument by the probability distribution  $P(z)$ . The model that we use is therefore:

$$P(f, t) = \sum_z P(z) \sum_{\tau_f} P_K(\tau_f|z) P_I(f - \tau_f, t|z)$$

Given the mixture constant-Q transform, we use a variation of the EM algorithm to estimate the kernel distribution and impulse distribution for each instrument, effectively performing a deconvolution.

We use the impulse distribution of each instrument to estimate the pitch track of that instrument. At each time step the frequency at which the impulse distribution has a maximum value is considered to be the fundamental frequency of the note played by the given instrument. It should be noted that a frequency shift in the impulse distribution in one direction can be cancelled by a frequency shift in the kernel distribution in the other direction. Due to this ambiguity, we obtain relative pitch tracks for each instrument.

## 4. SPARSE IMPULSE DISTRIBUTIONS

The impulse distribution of a given instrument,  $P_I(f', t|z)$ , should ideally be an impulse at each time step (frame) so that we can consider the peak at each time step to be the relative pitch of the given instrument. This implies a sparse impulse distribution. In a probability distribution, sparsity can be expressed by low entropy. We therefore use entropic prior dis-

tributions on the impulse distributions forcing them to have a low entropy. The entropic prior is expressed as:

$$P(\theta_{f',t|z}) \propto e^{-\beta \mathcal{H}(\theta_{f',t|z})}$$

Where  $\mathcal{H}(\theta_{f',t|z})$  is the entropy of the given impulse distribution and  $\beta$  is the strength of the prior distribution. Another advantage of having impulse distributions with a low entropy is that more information is pushed into the kernel distributions which summarize the spectral signature of each instrument. We would like all of the information in the constant-Q transform except for the pitch tracks to be explained by the kernel distributions. The entropic prior helps with this.

## 5. INSTRUMENT SPECIFIC PRIORS

The kernel distribution for each instrument captures the spectral signature of the given instrument. There are results in the musical acoustics literature that tell us about the nature of the spectra of specific instruments. We can use these results to bias the kernel distribution estimates by using different prior distributions for each instrument. The kernel distribution for a given instrument,  $P_K(\tau_f|z)$  is a multinomial distribution. Since the Dirichlet distribution is conjugate to the multinomial distribution, it is a good choice for a prior distribution. We specify our intuition of the spectrum of a given instrument as hyperparameters,  $\alpha(\tau_f|z)$  of the corresponding Dirichlet distribution. The prior distribution for a given kernel distribution would therefore be specified as:

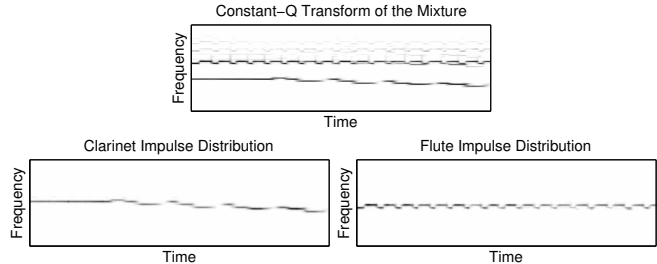
$$P(\Lambda_{\tau_f|z}) \propto \prod_{\tau_f} P_K(\tau_f|z)^{\kappa \alpha(\tau_f|z)}$$

We assume that  $\sum_{\tau_f} \alpha(\tau_f|z) = 1$  without any loss of generality.  $\kappa$  will therefore give us the strength of the prior distribution. An example of the use of such a prior is for specifying our intuition of the spectrum of a clarinet. From the musical acoustics literature, we know that the clarinet has missing even harmonics. We enforce this by using a harmonic series with missing even harmonics as the hyperparameters for the Dirichlet prior distribution on the kernel distribution that corresponds to the clarinet.

## 6. IMPULSE DISTRIBUTION SMOOTHING

The pitch tracks for each instrument are extracted from the estimated impulse distributions. Ideally, each impulse distribution would contain the complete pitch track for exactly one instrument and nothing else. In practice, the estimation process sometimes yields a switching of pitch tracks. This is to say that the pitch tracks of multiple instruments show up in each impulse distribution.

This issue is dealt with by imposing a temporal continuity constraint on each impulse distribution. We use a Kalman filter type smoothing for this purpose. As the EM algorithm is



**Fig. 1.** Illustration of the multipitch estimation on a mixture of a clarinet and flute. Shift invariant PLCA is performed on the mixture data that is shown in the top figure. The two bottom figures show the estimated impulse distributions for the two instruments. The pitch tracks can clearly be seen to correspond to the fundamental frequencies of the instruments.

an iterative algorithm, we obtain an estimate for each impulse distribution in the  $M$  step of each iteration. This estimate is denoted as  $P_I^*(f', t|z)$ . We apply our smoothing on this distribution at every EM iteration. This is done by multiplying the distribution at each time step by a Gaussian whose mean is the peak at the previous time step. Our smoothed impulse distribution at a given iteration is therefore:

$$P_{I_{smooth}}^*(f', t|z) = P_I^*(f', t|z) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(f' - \mu_{t'-1|z})^2}{2\sigma^2}}$$

The smoothed distribution,  $P_{I_{smooth}}^*(f', t|z)$ , is reassigned to  $P_I^*(f', t|z)$  before continuing with the next EM iteration.

## 7. RESULTS AND CONCLUSIONS

Initial experiments have shown results with accuracy greater than 90% on certain mixtures of two instruments. An example of the results can be seen in Fig.1. Further refinement of the entropic prior and instrument specific priors has the potential to greatly improve performance and robustness on a variety of sound mixtures.

We have presented an algorithm to concurrently estimate the pitch of multiple instruments in a sound mixture. A constant-Q transform of the input sound is deconvolved using shift-invariant probabilistic latent component analysis. Sparsity on the impulse distributions is achieved by using entropic prior distributions on the impulse distributions. This helps push most of the information apart from the pitch information into the kernel distribution. Kernel distributions are biased to represent the spectra of certain instruments by using prior distributions based on musical acoustics. Temporal continuity of the pitch tracks is enforced by using a Kalman filter type smoothing on the impulse distributions. In future work, we plan on exploring the use of other methods to achieve sparse impulse distributions. We also plan on exploring the use of more prior distributions to improve the modeling.