RELATIVE PITCH ESTIMATION OF MULTIPLE INSTRUMENTS

Gautham J. Mysore[*]

Center for Computer Research in Music and Acoustics
Stanford University

Paris Smaragdis

Advanced Technology Labs
Adobe Systems Inc.

ABSTRACT

We present an algorithm based on probabilistic latent component analysis and employ it for relative pitch estimation of multiple instruments in polyphonic music. A multilayered positive deconvolution is performed concurrently on mixture constant-Q transforms to obtain a relative pitch track and timbral signature for each instrument. Initial experimental results on mixtures of two instruments are quite promising and show high levels of accuracy.

***Index Terms*—** Pitch estimation, Sound mixtures, Unsupervised learning

1. INTRODUCTION

Pitch estimation of concurrent multiple instruments is an ongoing pursuit in the world of musical signal processing. Although the problem of pitch estimation of a single instrument is for most practical reasons a relatively easy problem to solve, when confronted with mixtures of instruments, monophonic approaches are ill-equiped to estimate multiple pitch values. This problem has been attacked using various methods based on auditory scene analysis [1], auditory models [2], Bayesian inference and model-based analysis [3, 4], and also by employing source separation followed by monophonic pitch estimation on the separated outputs. In this paper, we propose a different approach to the problem in which we use an unsupervised method that can elegantly deal with sound mixtures as well as with monophonic inputs.

Our approach operates on a constant-Q transform representation, which is decomposed by a deconvolution algorithm designed to find consistent spectral patterns and infer their pitch by observing their instantaneous shifts along the frequency axis. In order to make the estimation robust for mixtures, we use a parameterized model of the shifting process as well as a Kalman filter type smoothing to enforce temporal continuity in the extracted pitch tracks.

2. PROPOSED METHOD

In this section, we describe the various computational steps involved in constructing our approach. In the process, we also present some simple examples that justify and highlight the utility of these steps.

[*]This work was performed while at Adobe Systems.

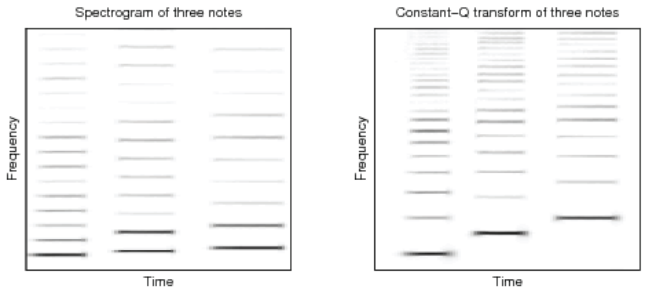

Fig. 1. Spectrograms and constant-Q transforms. The figure on the left is a spectrogram of three different notes played by a saxophone. The spacing between any two harmonics in the first note is different from the spacing between the two corresponding harmonics in the other notes. The figure on the right is a constant-Q transform of the same signal. The spacing between the harmonics is the same for all of the notes. In this representation a pitch shift is characterized by just a vertical shift of the frequency axis.

2.1. Constant-Q Transform

The initial representation we use for this method is the constant-Q transform [5]. The constant-Q transform is a time-frequency representation with a logarithmically spaced frequency axis. Due to this frequency spacing arrangement, when analyzing pitched signals, we can see pitch shifts represented as vertical shifts of roughly the same spectral template. This is different from a representation like the Fourier decomposition, which will additionally warp the spectral shape of an instrument as its pitch changes. This crucial difference is shown in the following example. In figure 1, we show a Fourier representation of a musical note in which one can clearly see its harmonic structure. If we visualize a different note of the same instrument, the spacing between the harmonics change. On the other hand, in a constant-Q transform, notes at different pitches appear as shifted versions of the same spectral pattern. If the timbral structure of a given instrument is fairly consistent in a piece of music, the constant-

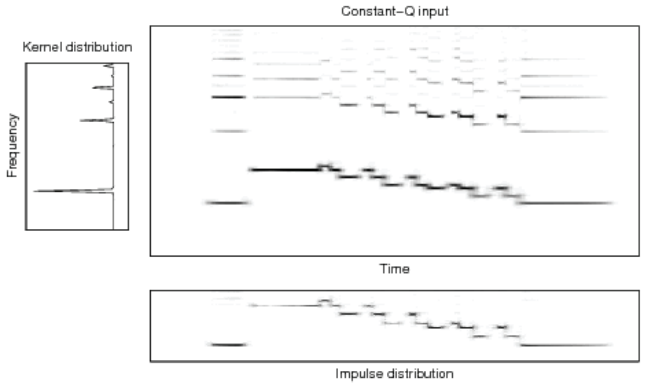

Fig. 2. Illustration of shift invariant PLCA on a recording of a clarinet. The large top right figure is the constant-Q transform of a clarinet recording. The bottom figure shows the impulse distribution after applying shift invariant PLCA and the top left figure shows the corresponding kernel distribution. Convolving the kernel and the impulse distribution will approximately reconstruct the input constant-Q transform.

Q transform of the instrument can be seen as a convolution of the spectral pattern of that instrument with a function that offsets it appropriately to create the desired pitch effects.

2.2. Shift-Invariant Probabilistic Latent Component Analysis

Shift-invariant probabilistic latent component analysis (PLCA) [6] is an algorithm used to extract shifted structure in multi-dimensional non-negative data. When employed on a constant-Q transform of a mixture of instruments, it can be used to decompose the input data into a summation of convolutions of one spectrum and one pitch track for each instrument.

More specifically, we denote the spectral signature of the z-th instrument as a probability distribution we call the kernel distribution $P_K(\tau_f|z)$ and we define the pitch track of the same instrument as a probability distribution we call the impulse distribution $P_I(f', t|z)$. The constant-Q transform of the given instrument is therefore the convolution of these two distributions (figure 2) :

$$V_{ft|z} = P_K(\tau_f|z) * P_I(f', t|z)$$

When we have a mixture, we will observe one $V_{ft|z}$ for each instrument and all of these will be superimposed to construct the constant-Q input at hand. We model their mixing proportion as one more probability distribution $P(z)$.

The model for the constant-Q transform of the mixture is therefore:

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

Since there are latent variables in this model, a variant of the EM algorithm is employed to estimate the distributions. The latent variables are τ_f (or f' since $f = \tau_f + f'$) which represents shift and z which represents the mixture weights.

In the expectation step, we estimate the contribution of a specific location of the impulse distribution (f',t) of a given instrument z, to location (f, t) of the mixture constant-Q transform:

$$R(f, t, f', z) = \frac{P(z) P_I(f', t|z) P_K(f - f'|z)}{\sum_z P(z) \sum_{f'} P_I(f', t|z) P_K(f - f'|z)}$$

In the above equation, the latent variables are f' and z. We can obtain the same value of this function by modeling the latent variables as τ_f and z. The E-step equation then becomes:

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

The M-step equations are given by the following update equations:

$$P_I^*(f', t|z) = \frac{\sum_f V_{ft} R(f, t, f', z)}{\sum_{f'} \sum_t \sum_f V_{ft} R(f, t, f', z)}$$

$$P_K^*(\tau_f|z) = \frac{\sum_f \sum_t V_{ft} R(f, t, \tau_f, z)}{\sum_{\tau_f} \sum_f \sum_t V_{ft} R(f, t, \tau_f, z)}$$

$$P^*(z) = \frac{\sum_{f'} \sum_t \sum_f V_{ft} R(f, t, f', z)}{\sum_z \sum_{f'} \sum_t \sum_f V_{ft} R(f, t, f', z)}$$

The above equations are iterated until convergence.

A number of different impulse/kernel distribution decompositions can combine to give the same constant-Q transform. A vertical shift of one distribution towards one direction can be annulled by a vertical shift of the other distribution towards the other direction. Due to this uncertainty, we can only extract a "relative pitch" track as opposed to an absolute pitch measurement. Of course, with due precessing it is easy to align that track and obtain the absolute pitch if needed.

Ideally, the impulse distribution at each time step would be a shifted and scaled delta function. The position of its peak would then give us the pitch at a given time step. Since there is some amount of averaging involved in the estimation process, the impulse distribution is smoother than an impulse at each time step. If this distribution is unimodal at each time step, we can just estimate the pitch track as its peak. This is however not always the case since the estimation is not required to be well-behaved like that. In order to ensure that each time step of the impulse distribution is a unimodal distribution with a clear peak value we employ a sliding-Gaussian Dirichlet prior distribution as described in the next section.

2.3. Sliding-Gaussian Dirichlet Prior

In order to deal with the potential non-unimodal nature of the impulse distribution at each time step, we use a "sliding

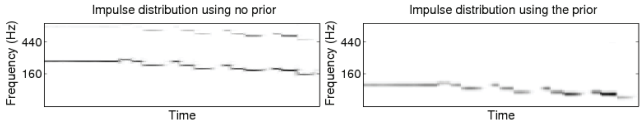

Fig. 3. Illustration of the use of the sliding-Gaussian Dirichlet prior. Impulse distributions without the use of the prior (left) and with the use of the prior (right). It can be seen that a harmonic is captured in the impulse distribution when no prior is used. The distribution becomes unimodal when the prior is used.

Gaussian" Dirichlet prior distribution. The prior distribution is used in the estimation of the impulse distribution in the M-step of each iteration, making this a maximum a posteriori (MAP) estimation.

We use the prior distribution to impose a prior belief that the impulse distribution of each instrument is unimodal at each time step, thus exhibiting a clear peak which we can interpret as the pitch value at that time. The effect of using this prior can be seen in figure 3. The hyperparameters of the Dirichlet prior at each time step therefore form a sampled and scaled Gaussian.

The hyperparameters are therefore defined as follows:

$$\alpha(f,'t|z) \quad = \quad \rho\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left(f'-\mu_{t|z}\right)^2}{2\sigma^2}}$$

where ρ is a parameter that allows us to decide the strength of the prior.

The prior distribution for the impulse distribution of the z-th instrument would then be:

$$P(\Lambda|z) \quad = \quad \frac{1}{\beta}\prod_{f',t}P_I(f',t|z)^{\alpha(f',t|z)}$$

where β is a normalizing factor. The M-step equation to estimate the impulse distribution then becomes:

$$P_I^*(f',t|z) = \frac{\sum_f\left(V_{ft}R(f,t,f',z)+\rho'\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left(f'-\mu_{t|z}\right)^2}{2\sigma^2}}\right)}{\sum_{f'}\sum_t\sum_f\left(V_{ft}R(f,t,f',z)+\rho'\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left(f'-\mu_{t|z}\right)^2}{2\sigma^2}}\right)}$$

where ρ'= $\frac{\rho}{\beta}$. It can be seen from the numerator of this equation that at each time step, we are performing a blend of the previous (using no prior) unnormalized estimate of the impulse distribution and a Gaussian. The variance of the Gaussians σ^2, is predetermined. The peak of the previous unnormalized estimate of the impulse distribution is used as the mean at each time step.

$$\mu_{t|z} \quad = \quad \arg\max_{f'}\sum_f V_{ft}R(f,t,f',z)$$

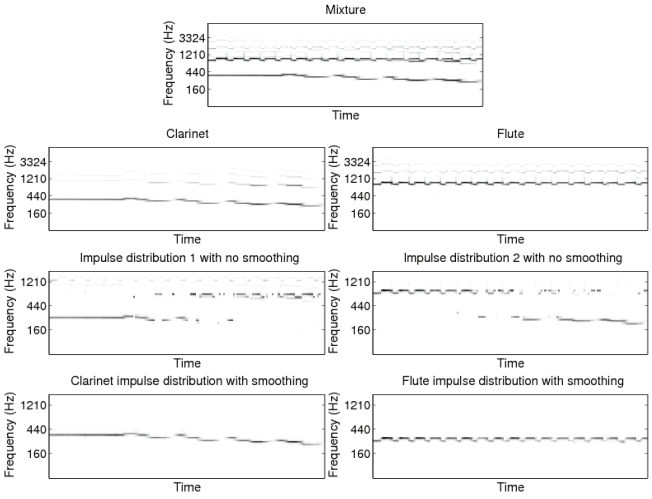

Fig. 4. Illustration of the effect of impulse distribution smoothing on multiple instruments. Shift invariant PLCA is performed on the mixture data that is shown in the top figure. The two figures in the second row are constant-Q transforms of the individual instruments that combine to form the mixture. The two figures in the third row show the resulting impulse distributions when temporal smoothing is not used. As can be seen, both impulse distributions contain elements of both instruments. The bottom two figures show the resulting impulse distributions with the use of temporal smoothing.

2.4. Impulse Distribution Smoothing

The goal of our estimation is to obtain a separate relative pitch track for each instrument. Sometimes however, there is a temporary switch between pitch tracks of different instruments in a given impulse distribution. We can see that in figure 4, where a given impulse distribution oscillates between two different instrument pitch tracks. At a given time step, the impulse distribution is predominantly unimodal (with the help of the prior). However, there are constant oscillations between pitch tracks as can be seen in the figure.

In order to deal with this issue, we impose a temporal continuity constraint on each impulse distribution so that large variations between consecutive time steps are discouraged. This is done by employing a Kalman filter type smoothing. This is done by multiplying the impulse distribution at each time step with a Gaussian whose mean is the peak of the impulse distribution at the previous time step:

$$P^*_{I_{smooth}}(f',t|z) \quad = \quad P^*_I(f',t|z)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\left(f'-\mu_{t'-1|z}\right)^2}{2\sigma^2}}$$

The variance σ^2 is predetermined. Once we obtain $P^*_{I_{smooth}}(f',t|z)$, we reassign it to $P^*_I(f',t|z)$ and continue with the next EM iteration.

Clarinet and Flute	μ_1	σ_1	μ_2	σ_2
Metric 1 Error (percent)	2.22	7.61	1.06	5.41
Metric 2 Error (bin #)	0.15	0.9	0.21	1.54
Flute and Horn	μ_1	σ_1	μ_2	σ_2
Metric 1 Error (percent)	4.12	7.48	6.89	15.82
Metric 2 Error (bin #)	0.25	0.83	0.45	3.2
Clarinet and Oboe	μ_1	σ_1	μ_2	σ_2
Metric 1 Error (percent)	26.91	13.28	10.74	9.69
Metric 2 Error (bin #)	2.58	4.28	1.92	5.85

Table 1. Results. The mean and standard deviation of the error over each metric is found for each instrument. For a given error metric, the mean and standard deviation for each instrument are indicated by their subscripts.

3. RESULTS

Our algorithm has been tested on a recording of a woodwind quintet [1]. We have applied the algorithm to mixtures of clips of two instruments at a time. Since the EM algorithm does not always converge to the same solution, we have run the algorithm for one hundred trials on each of the three data sets (mixture of two instruments) that we have used.

The ground truth is obtained by finding the position of the peak of the constant-Q transforms of solo instruments (such as the second row of figure 4) at each time step (frame). We first align the relative pitch track obtained in each of the trials with this ground truth data. As can be seen in figure 5, the majority of the trials (using the first mixture) converge to the same correct solution. Ninety-two out of these one hundred trials actually converge to an almost identical solution.

We then compute two error metrics. For the first metric, we find the percentage of misclassified frames in each of the trials. A frame is considered to be misclassified if the estimated pitch differs from the ground truth by more than one constant-Q bin as this corresponds to half a semitone (using a constant-Q transform with 24 bins/octave). We have computed the mean and the standard deviation of this error percentage for each instrument over the one hundred trials.

For the second metric, we find the difference between the estimated pitch and the ground truth (in number of constant-Q bins) at each frame. We then find the mean and standard deviation of the number of bins over all one hundred trials.

As seen in table 1, the results are very satisfactory for the first two mixtures. The error using the first metric is between 2.22% and 6.89%. The error using the second metric is always less than 1 bin. If we use the best ninety-two trials for the first mixture, the errors go down to less than 1% and less than 0.1 bin. The errors are higher in the third mixture. However, the error using the second metric is still less than 3 bins.

[1] From the development set for the MIREX 2007 multiF0 estimation tracking task.

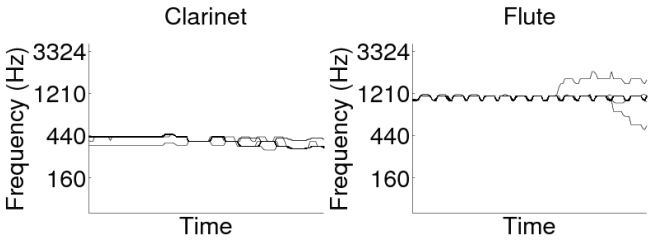

Fig. 5. Overlay plots of aligned pitch tracks. The resulting pitch tracks from one hundred trials of our algorithm on a mixture of a clarinet and a flute have been aligned and overlayed. As can be seen, the majority of the trials converge to the same correct solution.

4. CONCLUSIONS

We have presented an unsupervised learning algorithm that is used for the estimation of the pitch of multiple concurrent instruments and have demonstrated the algorithm on mixtures of two instruments. The use of a prior distribution and temporal smoothing has been shown to improve certain shortcomings of the algorithm. This method is quite promising and in future work, we plan to improve the performance by modeling musical structure.

5. REFERENCES

[1] D. Ellis, *Prediction-Driven Computational Auditory Scene Analysis*. PhD thesis, M.I.T., 1996.

[2] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, pp. 255–266, February 2008.

[3] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[4] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 679–694, March 2006.

[5] J. C. Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, January 1991.

[6] P. Smaragdis, B. Raj, and M. V. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proceedings IEEE International Conference on Audio and Speech Signal Processing*, April 2008.