Relative-Pitch Tracking of Multiple Arbitrary Sounds

Paris Smaragdis[a)]
Adobe Systems Inc., 275 Grove St. Newton, MA 02466, USA

Perceived-pitch tracking of potentially aperiodic sounds, as well as pitch tracking of multiple simultaneous sources is shown to be feasible using a probabilistic methodology. The use of a shift-invariant representation in the constant-Q domain allows the modeling of perceived pitch changes as vertical shifts of spectra. This enables the tracking of these changes in sounds with an arbitrary spectral profile, even those where pitch would be an ill-defined quantity. It is further possible to extend this approach to a mixture model which allows simultaneous tracking of varying mixed sounds. Demonstrations on real recordings highlight the robustness of such a model under various adverse conditions, and also show some of its unique conceptual differences when compared to traditional pitch tracking approaches.

PACS numbers: 43.66.Jh, 43.75.Zz

I. INTRODUCTION

Pitch tracking has long been been a fascinating subject in musical acoustics. This is a problem which has been tackled using a rich variety of approaches and continues to inspire a considerable amount of research. Approaches to pitch track extraction have ranged from straightforward period estimation, to sophisticated statistical methods, some employing time domain techniques and others sophisticated front-ends that reveal more of the pitch structure [1–8]. The applications of pitch tracking cover a wide range of applications ranging from musical transcription, to emotion recognition in speech, to animal acoustics. To facilitate such a wide variety of applications various biases are often imposed to facilitate a reasonable answer for the domain at hand. In this paper we present a general approach to tracking a pitch-like measure which makes minimal assumptions about the nature of the input sound, or the kind of pitch content at hand.

We present an additive shift-invariant decomposition which when coupled with a constant-Q analysis front-end can be used to track movements of spectral structures along the log-frequency space. Such movements correlate very strongly to how we perceive pitch, and can be used to infer relative-pitch changes. Using this model we set forth to address a number of issues. The primary goal is to present a formulation which allows *soft* decisions which do not result in deterministic estimates, but rather a probability distribution describing the relative likelihood of these shifts along the frequency axis. This probabilistic approach, which is most valuable when designing systems with input of high uncertainty, also provides an easy way to extend such a system by using statistical methods that take advantage of domain knowledge that can further help achieve robust performance. An additional point we wish to address is that of tracking in the case of mixtures. The assumption of clean input sounds is rarely valid in real recordings, and often we need to compute pitch tracks of either noisy or multiple sources. The model we present is additive by design, so that multiple overlapping frequency shifts can be tracked simultaneously. This allows us to process inputs with multiple sources without serious complications. Finally, using this particular representation allows us to deal with unpitched or inharmonic sounds whose absolute pitch value is hard to pinpoint, yet they can be used to a melodic effect. Examples of such cases are chords, or certain percussive instrument sounds (e.g. cymbals) that individually do not have a strong pitch characteristic, but once used in a certain succession they can invoke the percept of a melody, or tonality. In these cases, the tracked shifts along the frequency axis provide an indication of a likely perceived melody, something that methods based on harmonicity assumptions would not be able to provide. Through various experiments we show that this approach can deal with sound sources which have challenging spectral profiles, as well as sources that exhibit a dynamic spectral character,

The remainder of this paper is structured as follows. We will begin by describing a frequency shifting approach to modeling pitch changes, we briefly discuss the constant-Q transform and its utility for our purposes, we will then introduce the computational details of our approach, and then demonstrate it with a variety of pitch tracking situations that highlight its abilities to overcome difficult situations.

II. A SPECTRAL SHIFTING APPROACH TO MODELING PITCH CHANGES

Pitch, especially as we perceive it, is an elusive concept. Most trivially we can link it to the fundamental vibrating frequency of a sound-generating object. However it is very easy to find examples of aperiodic, or otherwise harmonically incoherent sounds where this assumption can break. Because pitch is hard to estimate, and in some cases non-existent, attempting to construct pitch tracks in terms of a series of instantaneous pitches is an inherently risky endeavor. Instead, in this paper, we use a different approach which we argue is more akin to how we perceive pitch.

We will approach the pitch tracking problem as a fre-

[a)] Electronic address: paris@adobe.com

quency shift problem. Instead of trying to estimate absolute pitch quantities at every point in time, we will instead track relative changes of pitch across time. This is very similar to how most of us perceive pitch where we can note relative changes but not necessarily actual values. Aside from this connection, more importantly we sidestep the issue of defining and estimating the actual pitch. Instead of pitch measurements, we track the movement of spectral structures along the log-frequency axis. These shifts correlate very much to our perception of a pitch change and can be directly used to infer pitch movements. This can also allow us to deal with inharmonic or aperiodic sounds which in isolation do not have a clearly defined pitch, but when modulated create that percept. To accommodate this broader notion of pitch we will be using the term *spectral pitch* to indicate this particular movement across the frequency axis.

In the next two sections we will describe the representation that can reveal this modulation, and the machinery involved in detecting it.

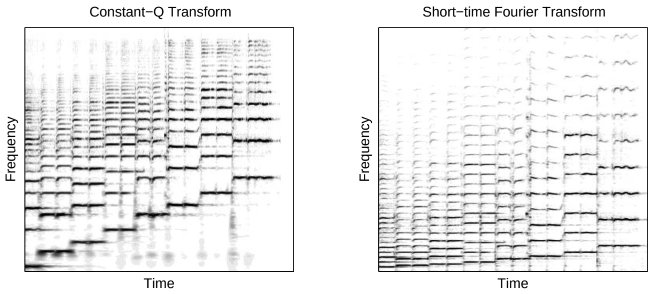


FIG. 1. A comparison between the constant-Q and the short-time Fourier transform. The input is a recording of an arpeggio performed by a violin. Note how in the constant-Q transform shown on the left the individual harmonics of the violin sound maintain their relative distance regardless of the note being played, whereas in the short-time Fourier transform they get spread apart as the notes move to higher frequencies.

III. CONSTANT-Q REPRESENTATIONS OF SOUNDS

The constant-Q transform is a time/frequency decomposition that exhibits a logarithmic frequency scale [9]. It is defined so that each octave of available frequencies spans the same number of frequency bins. Examining the magnitude of the constant-Q transforms results in visualizing the amount of acoustic energy at any point in the time/frequency plane. For the remainder of this paper we will be referring to the magnitude of the constant-Q transform and discard the phase information.

A very important property of this type of transformation is that changes in spectral pitch can be clearly visualized as shifts along the frequency axis. An example of this, as contrasted with the short-time Fourier transform, is shown in figure 1. On the left we show the constant-Q transform of an arpeggio performed on a real violin, and on the right its equivalent through a short-time Fourier transform. Upon closer examination it is easy to see that the note changes in the constant-Q plot are represented as vertical shifts of approximately the same spectral shape. For the short-time Fourier transform the spacing between the individual harmonics becomes wider for higher notes. This observation will be our starting point in defining the tracking model in this paper. Noting that in the constant-Q transform, the major variation that distinguishes different notes of the same instrument is a simple shift along the frequency axis, we will endeavor to track it and interpret it as a pitch movement.

An underlying assumption in this model is that the spectral shape of an individual sound is relatively constant as it changes pitch, so that the measurement of the shift is feasible. Theoretical arguments on that point are difficult to make since they rely on the expected statistics on the inputs, but as we will demonstrate later on this assumption holds well for sounds with widely varying spectral character.

Another point we need to make here is that of the approximate additivity of the magnitude constant-Q transform. The actual transform results in a complex valued output and is a linear operation which maintains that the transform of the sum of two signals equals the sum of the transforms of the two signals. When we compute the magnitude of the transform however there is no guarantee of linearity since for pair of any complex numbers $\{z_1, z_2\} \in \mathbb{C}$ we have $||z_1|| + ||z_2|| \neq ||z_1 + z_2||$. However when observing mixtures of multiple sounds there is often a high degree of disjointness in their spectra and the likelihood of both sounds being significantly active at the same time/frequency cell is often very low. In addition to that we seldom observe complete phase cancellations so even in the cases where significant energy overlaps we still have an effect approximate to addition. This assumption has been very commonly used for multiple audio processing systems and is generally understood to be valid for practical purposes. Under this assumption, when we observe the mixture of multiple notes we will expect the observed constant-Q transform to be composed out of the addition of constant-Q transforms that are appropriately composed out of shifted spectra, denoting each instrument or each note being played. This complicates the operation we wish to resolve, by requiring that we track potentially multiple spectra, that shift independently. If the input is composed of the same sound exhibiting multiple simultaneous pitches (such a polyphonic piano passage), then we would observe the same spectrum being shifted and overlaid accordingly for each note. If we have multiple instruments we would expect each instrument to have its own spectral shape which shifts and overlays according to its melodies. In the following section we will present an algorithm that allows us to track these simultaneous shifting movements, and help us interpret them as a relative spectral pitch change.

IV. A MODEL FOR TRACKING SPECTRAL PITCH SHIFTS

The computational model we will use in this section is the one developed in [10]. For reasons which will become clearer later on, we will be interpreting the magnitude constant-Q transform as a probability distribution. The contents at the time and frequency coordinates t, ω will be interpreted as an arbitrary scaling of the probability of existence of energy at that point. Due to this we will notate constant-Q transforms as $P(\omega, t)$ and assume the proper scaling so that they integrate to unity.

A. The single source formulation

Starting with this model we wish to discover shift invariant components across the ω dimension. In the simple case where we assume one shifting spectrum we can notate this model as:

$$P(\omega, t) = P_K(\omega) * P_I(f_0, t) \tag{1}$$

where $P(\omega, t)$ is the input magnitude constant-Q transform, $P_K(\omega)$ is a frequency distribution, $P_I(f_0, t)$ is a time/frequency distribution and the star operator denotes convolution (note that the convolution is two dimensional since the second operant is of that rank). We will refer to $P_K(\omega)$ as the *kernel distribution*, and $P_I(f_0, t)$ as the *impulse distribution*. Their interpretation is rather straightforward. The kernel distribution $P_K(\omega)$ is a frequency distribution, i.e., a constant-Q spectrum, or rather a prototypical vertical slice of a constant-Q transform. The impulse distribution $P_I(f_0, t)$ is a time/frequency distribution which gets convolved with the kernel distribution. Due to this relationship we can interpret the impulse distribution as expressing the likelihood that the kernel distribution will take place at any given frequency shift or time. To illustrate this concept consider the case in figure 2. In the top right panel we show the constant-Q transform of a recording of a real violin performing a glissando with vibrato. The ideal decomposition given our model is also shown. To the left we see the kernel distribution which denotes the spectral character of the violin, and in the bottom plot we see the corresponding impulse distribution. Convolving these two distributions we would approximate the input. It is easy to see that the impulse distribution graphically represents the frequency shift variations in a very convenient manner. In fact if we assume that the present instrument is well defined by the kernel distribution we can interpret the impulse distribution as a probability distribution of frequency shift, and by extension spectral pitch, across time. Because we also learn the actual spectral character of the input sound, we do not impose any requirements that the source has to be harmonic or otherwise structured, as long as the spectral pitch change is characterized by a shift in the frequency axis. As we will demonstrate later on this allows us to deal with arbitrary sounds very easily.

At this point, this model is quite closely related to the one in^4, and any such similar approach that employs shift-tracking on a log frequency scale. The main points of difference between what we present, and these approaches can be summarized in two points. First, we do not make the assumption that the signal we track is harmonic. Unlike past work we do not assume a known harmonic template whose movements are being tracked. In our framework, we allow for the flexibility to learn the particular spectral structure that characterizes the sound we track, which, as we show later, can allow us to use this approach even in cases where what we track is not clearly defined as pitch. Secondly, such techniques traditionally use cross-correlation to find the most likely placement of a harmonic series along the log-frequecy axis. The most likely placement is taken to be the peak of the cross-correlation. In our approach we actually produce a distribution which describes the likelihood of shift, which is a much more informative measure, especially when it needs to be incorporated into a larger reasoning system. The implied non-negativity in this operation also means that we won't be obtaining negative cross-correlation values which can obscure the interpretation of this operation, and make use of cross-cancellations which can impede finding the true function peak. This last argument will become increasingly more important as we move into multi-source formulations in the following sections.

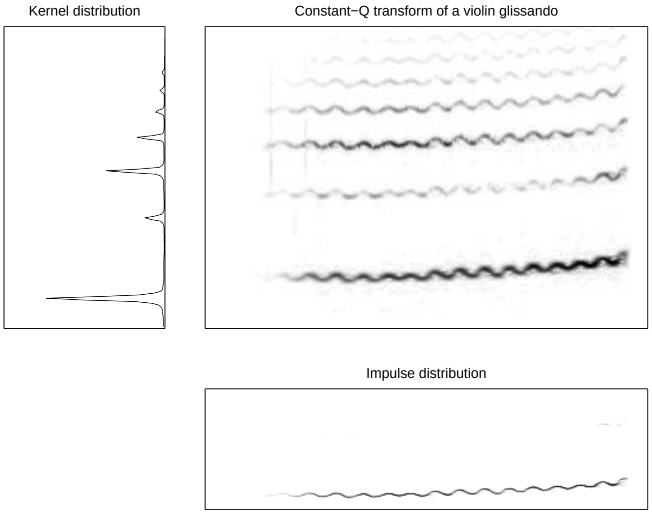


FIG. 2. An illustrative analysis of the constant-Q transform in the top right plot. The top left plot shows the extracted kernel distribution and the bottom plot shows the corresponding impulse distribution. The impulse distribution can be used to represent the spectral pitch changes in the input.

B. The multi-source formulation

In the case where we expect to encounter multiple sounds with different spectra, we can generalize the above model to:

$$P(\omega, t) = \sum_{z=1}^{R} P(z) P_K(\omega|z) * P_I(f_0, t|z) \tag{2}$$

The difference with equation 1 is that now we use a latent variable z as an index to allow us having R distinct kernel distributions $P_K(\omega|z)$, each with its own corresponding shifting pattern denoted by R impulse distributions $P_I(f_0, t|z)$. We also introduce a prior distribution $P(z)$, which allows us to arbitrarily weight these pairs of convolution operands to approximate the input. This is essentially an additive generalization of the previous model which allows the simultaneous tracking of spectral pitches by multiple sources with distinct spectral characters. Each kernel distribution conditional on z will describe the spectral shape of each source, and each impulse distribution conditional on z will describe its corresponding spectral pitch track. The priors distribution will effectively denote the mixing proportions of each spectral template, or more simply how much of it we will observe in relation to the others. The choice of R is up to the user. For $R = 1$ this model collapses to the model in the previous section, and tracks the movement of a single spectral template across the log-frequency space. If we know that the input we are analyzing is constructed by multiple and distinctly different spectra then we can set R to their count so that we can simultaneously track the shifts of multiple spectral templates at the same time.

C. Learning the model

In order to estimate the unknown distributions $P_K(\omega|z)$, $P_I(f_0, t|z)$ and $P(z)$ in the above model we can use the Expectation-Maximization algorithm [11]. We first rewrite the model in order to express the convolution in a more explicit manner as:

$$P(\omega, t) = \sum_{z=1}^{R} P(z) \sum_{f_0} P_K(\omega - f_0|z) P_I(f_0, t|z) \qquad (3)$$

EM estimation will break down the learning process in an iterative succession of two steps. The first step, the E-step, computes the contribution of each spectral template to the overall model reconstruction by:

$$Q(\omega, t, f_0, z) = \frac{P(z) P_K(\omega - f_0|z) P_I(f_0, t|z)}{\sum_{z'} P(z') \sum_{f_0'} P_K(\omega - f_0'|z') P_I(f_0', t|z')} \qquad (4)$$

This results in a "weighting" factor which tells us how important each kernel distribution is in reconstructing the input at any possible shift in the time/frequency space. During the second step, the M-step, we estimate the wanted model distributions by essentially performing matched filtering between each operant and the input weighted by the appropriate $Q(\omega, t, f_o, z)$. The equations for the M-step are:

$$\begin{array}{rcl} P(z)^* & = & \sum_\omega \sum_t \sum_{f_0} P(\omega, t) Q(\omega, t, f_0, z) \\ P_K(\omega|z)^* & = & \frac{\sum_t \sum_{f_0} P(\omega + f_0, t) Q(\omega + f_0, t, f_0, z)}{\sum_{\omega'} \sum_t \sum_{f_0} P(\omega' + f_0, t) Q(\omega' + f_0, t, f_0, z)} \\ P_I(f_0, t|z)^* & = & \frac{\sum_\omega P(\omega, t) Q(\omega, t, f_0, z)}{P(z)^*} \end{array} \qquad (5)$$

where the $(\cdot)^*$ notation indicates the new estimate. Iterating over the above steps we converge to a solution after about 30 to 50 iterations. Although there is no guarantee that this process will find the global optimum, it predominantly converges to qualitatively the same solutions over repeated runs. The dominant variation in these solutions is an arbitrary shift in the spectral distribution which is counteracted by an opposing shift in the impulse distribution in order to ensure the correct reconstruction. This introduces a variation in the resulting outputs, but not one that interferes with the quality of fit, or (as we examine in later sections) with the interpretation of the decomposition.

D. Sparsity constraints

Upon closer consideration one can see that the above model is overcomplete. This means that we can potentially have more information in the model itself than we have in the input. This is of course a major problem because it can result in outputs which have overfit to the input, or models which are hard to interpret. A particular instance of this problem can be explained using the commutativity property of convolution. Referring back to figure 2 we note that an output in which the impulse distribution was identical to the input and the kernel distribution was a delta function would be also an acceptable answer. That particular decomposition wouldn't offer any information at all since the spectral pitch track would have be identical to the input. Likewise any arbitrary shift of information from one distribution to another that would lie between what we have plotted above and the outcome just described, would result in an infinite set of correct solutions.

In order to regulate the potential increase of information from input to output we will make use of an *entropic prior* [12]. This prior takes the form of $P(\theta) \propto e^{-\beta \mathcal{H}(\theta)}$, where $\mathcal{H}(\cdot)$ is entropy and θ can be any distribution from the ones estimated in our model. The parameter β is simply adjusting the amount of bias towards a high or a low entropy preference. A β value which is less than zero will bias the estimation towards a high entropy (i.e., flatter) distribution, and a positive value will bias it towards a low entropy (i.e., spikier) distribution. The magnitude of β determines how important this entropic manipulation is so that larger values will put more stress in it, whereas values closer to 0 will not. In the extreme case where $\beta = 0$ the prior does not come in effect. Imposing this prior can be done by inserting an additional procedure in the M-step which enforces the use of this prior. The additional step involves re-estimating the distribution in hand by:

$$\theta^{**} = \frac{-\left(\hat{\theta}^*\right)/\beta}{\mathcal{W}(-\left(\hat{\theta}^*\right) e^{1+\lambda/\beta}/\beta)} \qquad (6)$$

where $\mathcal{W}(\cdot)$ is Lambert's function [13], θ^{**} is the estimate of θ with the entropic prior, and $\hat{\theta}^*$ is the estimate according to equations 5 but without the division by $P(z)^*$. The quantity λ comes from a Lagrangian due to the constraint that $\sum \theta_i = 1$ which results in the expression:

$$\lambda = -\left[\frac{\hat{\theta}^*}{\theta_i^{**}} + \beta + \beta \log \theta_i^{**}\right] \quad (7)$$

These two last equations can be repeatedly evaluated in succession and they converge to a stable estimate of θ after a small number of iterations. This prior and the computational details as they relate to this model involved are described in more detail in [14].

With the problem at hand we would want to have a high entropy kernel distribution and a low entropy impulse distribution. This will result in a frequency-shift track which will be as fine as possible and a spectrum estimate which will account for most of the energy of the source's spectrum. To illustrate the effect of this prior consider the different outcomes shown in figure 3. The input was the same as in figure 2. The top plots show the results we obtain using the entropic prior, whereas the bottom plots show the results we get if we don't use it. When using the entropic prior we bias the learning towards a high entropy kernel and a low entropy impulse. This resulted in a clean and sparse track, as opposed to the non-interpretable one when not using the prior.

For practical reasons we can strengthen the low entropy bias by requiring the height of the impulse distribution to be only as big as the expected pitch range (therefore implying that values outside that range will be zero thus lowering the entropy). This results in faster training requiring convolutions of smaller quantities, but also restricts the possible solutions to the range we are interested in, thus aiding in a speedier convergence. The convolutions involved can also be evaluated efficiently using the fast Fourier transform. Training for the examples in this paper took less than a minute on a current mid-range laptop computer using a MATLAB implementation of this algorithm.

E. Interpreting the model

Let us now examine exactly how the results of this analysis could be interpreted. As we have stressed before, this approach tracks frequency shifts which correspond to relative-pitch movements, and does not compute an absolute spectral pitch. This means that at any point in time we do not know what the spectral pitch is, but rather how much it has changed in relation to other parts. To illustrate this idea consider the plots in figure 4. Each row of plots displays the results from a different simulation on the input in figure 2. Note that although the results appear somewhat identical they still differ by an arbitrary vertical shift. This shift is counterbalanced between the kernel and the impulse distribution such that when they convolve they result in the same output. However, we cannot expect the impulse distribution between multiple runs to exhibit the same shift since the model is shift-invariant. This means that we can recover the relative pitch changes in the input, but we cannot infer the actual spectral pitch values. If one is inclined to mark the fundamental in the kernel distribution then we can easily obtain the spectral pitch values. However detecting the fundamental will not always be a trivial pursuit, especially when dealing with noisy or aperiodic kinds of sounds. Regardless, for the scope of this paper, our objective is to detect relative changes and not the actual spectral pitch so we refer to this estimation as something that can be pursued in future work. Another issue that this analysis presents is that of the relation-

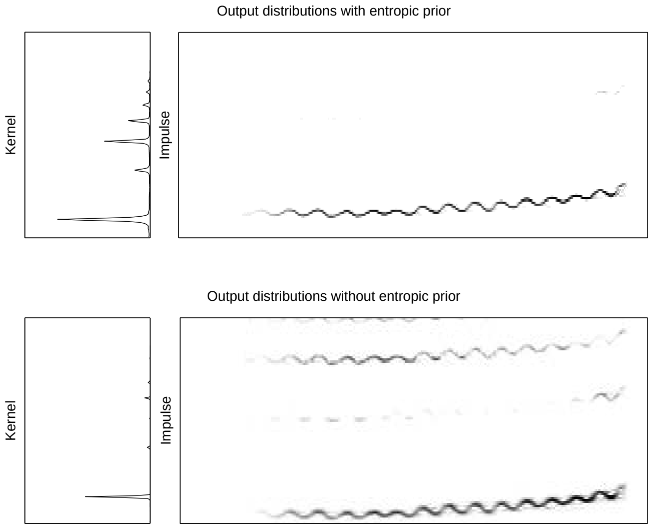


FIG. 3. Illustrating the effect of the entropic prior. The top figures show the output we obtain when analyzing the example in figure 2 and employ the entropic prior with a high entropy constraint on the kernel distribution and a low entropy constraint on the impulse distribution. The pair of the bottom plots show the kernel and impulse distributions learned without using the prior.

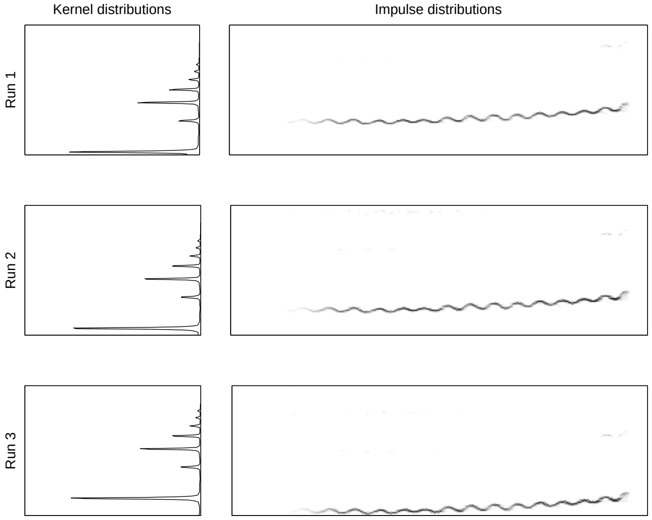


FIG. 4. Results from multiple runs on the input shown in figure 2. Notice that although in a qualitative sense the results are identical, there are vertical shifts between them that differentiate them. When the kernel distribution is positioned higher in the frequency axis, the impulse distribution is positioned lower and vice-versa.

ship of the output with the intensity of the input. As the intensity of the input fluctuates across time, we will see a similar fluctuation in the intensity of the impulse distribution across time. This is an effect that can be seen in all the previous plots, especially during the attack portion, where the impulse distribution transitions from faint to strong as it follows the performed crescendo. Additionally, at times where the violin is not performing we do not observe any intensity in the impulse distribution. This is because the impulse distribution is a distribution over both frequency and time, measuring the presence of the relative amount of presence of the kernel distribution as compared to all time points and frequency shifts. If we are only interested in the frequency offset of the kernel distribution then we need to examine each time slice of the impulse distribution. This is the distribution $P(f_0|t) = P(f_0,t)P(t)$, where $P(t)$ is overall input energy at time t, i.e. $P(t) = \int P(f_0,t)df_0$. To estimate the most likely frequency shift at time t we only need to find the mode of $P(f_0,t)$. However using $P(f_0,t)$ instead of $P(f_0|t)$ is a more intuitive choice since the estimate we will be normalized by the likelihood that the signal is active at that point in time, and thus also provide us with an amplitude estimate which we can use for note onset. If we decide to use $P(f_0|t)$ instead the silent sections will be quite uniform indicating that there is no dominant candidate for a frequency shift. This can be interpreted either as an unpitched section, or a silence. In order to avoid this ambiguity we perform the estimation on $P(f_0,t)$ which offers a more user friendly format.

V. EXAMPLES

Let us now show how this approach works with some more complex sounds, especially with challenging situations where conventional pitch tracking techniques can result in unreliable, or hard to interpret, estimates.

A. Single source, monophonic examples

In this case we will show results from analyzing two real sound samples containing only one note instance at a time. The first example is a violin recording performing a descending arpeggio with each note (except the first and the last) played twice in succession. The signal was analyzed from a frequency range of 300Hz to 8000Hz. The results of the analysis are shown in figure 5. The kernel distribution is clearly a harmonic series (on a constant-Q frequency scale), and the impulse distribution clearly shows the notes that were played. Subtle nuances such as the pitch correction at the beginning of the first note as well as the double note repeats can be easily seen. There are some artifacts as well, mostly in the form of octave echoes which are more present during the low notes. Since this representation displays the likelihood of spectral pitch these are not necessarily erroneous estimates since they are clearly of lower likelihood than the actual note, and represent the periodicity along the frequency axis of the constant-Q transform. Picking the maximal values of each column of the impulse distribution will easily result in the correct relative spectral pitch estimate at that time. Another more challenging example is shown in figure 6. In this case the analyzed source is a vocal recording of the first five notes of a major scale, each note being sung with a different vowel. This experiment is used to test the assumption that the spectrum of the input has to be constant. As is clearly seen in the figure the spectral character of each note is substantially different from the others. Examining the results we observe an averaged spectrum as the converged kernel distribution, and an appropriate frequency-shift track from the impulse distribution. It is important to stress this robustness when dealing with spectrally dynamic sounds since our original assumption of a constant spectrum is unrealistic for real recordings. It is well known that musical instruments exhibit a varying formant character at different registers and that not all notes can be modeled as a simple shift of others. As shown by all the experiments in this paper (and more so by the current one) the constant spectrum assumption in this approach, is not very strict at all and doesn't pose any serious problems with dynamically changing sources. For the final example of the monophonic cases we show how this approach performs when dealing with highly inharmonic sounds. The input in this case were the first four bars of the Deep Purple recording of the song "Smoke on the Water". The recording features a well known guitar pattern of an interval of a fifth being appropriately transposed to form a characteristic melody. The guitar sound is highly distorted which in addition to the fact that the melody involves multiple notes, creates a highly inharmonic sound which technically does not exhibit pitch (although perceptually it sounds tonal). However, since the same sound is being transposed to form a melody it is clearly perceived by a human listener as a melodic sequence. Figure 7 shows

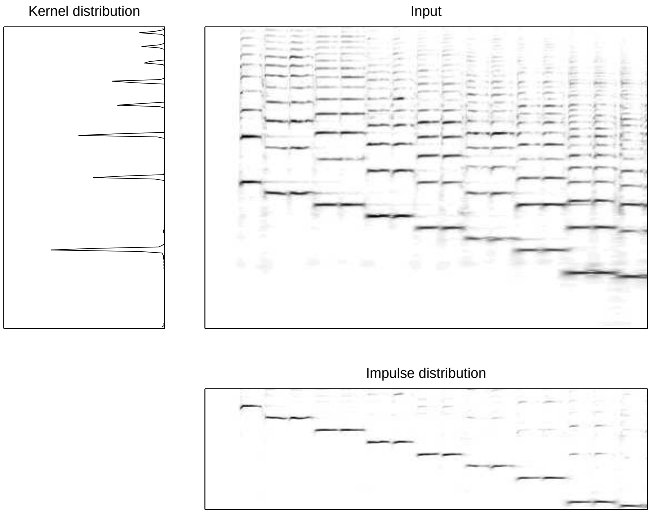
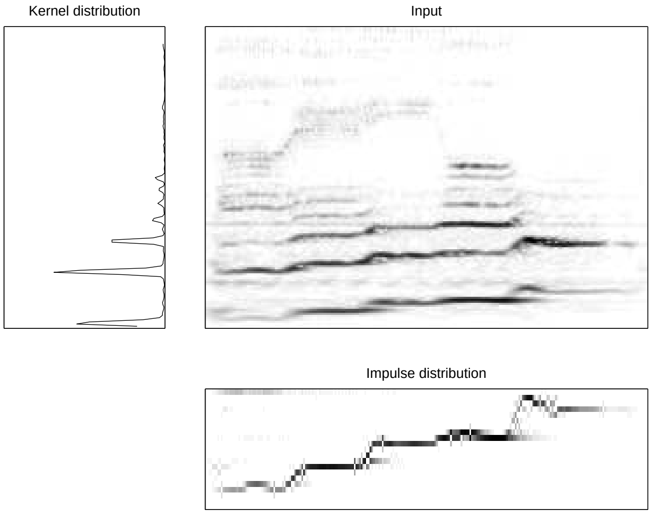


FIG. 5. Analysis of a violin arpeggio. The top right plot is the input constant-Q transform. The left plot is the extracted spectrum of the sound and the bottom plot is the estimated frequency shift track.

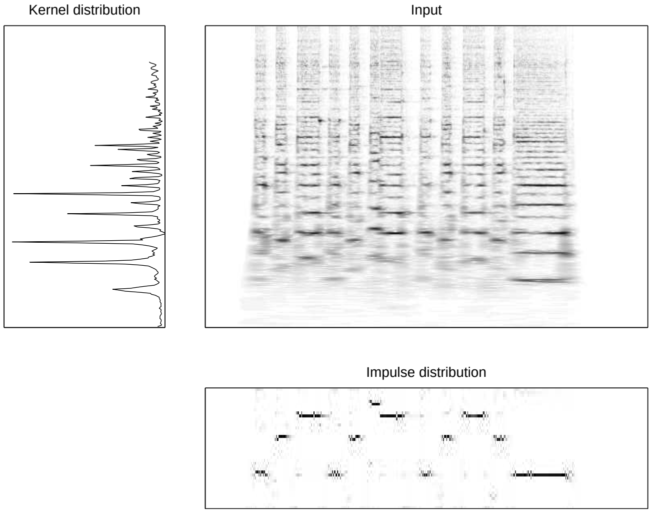


Kernel distribution

Input

Impulse distribution

FIG. 6. Example of spectral pitch tracking singing voice with a changing spectral character. The left plot is the extracted spectrum of the input and the bottom plot is the implied spectral pitch track.



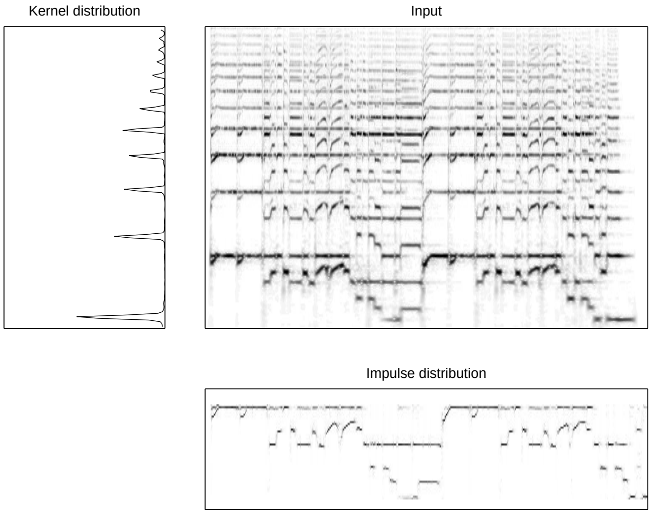
Kernel distribution

Input

Impulse distribution

FIG. 7. Analysis of the first four bars of Deep Purple's "Smoke on the Water". Despite the absence of a strict pitch at any point in time the transposed chord sequence forms a melody which we can clearly represent in the impulse distribution.

the results of this analysis. The melodic line is clearly displayed in the impulse distribution, and the spectral profile of the source as represented by the kernel distribution is as expected a highly inharmonic and busy spectral pattern. Similar results can be obtained when using inharmonic or aperiodic sounds, such as cymbals, tom-toms or bells, without any complications due to their non-harmonic spectral character.



Kernel distribution

Input

Impulse distribution

FIG. 8. Analysis of a country violin recording which involves the continuous simultaneous sounding of two notes. The impulse distribution clearly represents the two notes at any point in time and provides an accurate description of the melodic content.

B. Single source, multiple notes

Since the model is additive it is also able to deal with multiple notes. An example of this case is shown in figure 8. The input in this case was a country violin recording which included the simultaneous playing of two notes during most of the time. As expected the analysis of this sound results in an impulse distribution which has multiple peaks at each time frame that represent the two notes sounding at that point. In the impulse distribution it is easy to see the long sustained notes and the simultaneous ornamental fiddle playing.

C. Multiple sources, multiple notes

Finally we demonstrate the ability of this approach to deal with multiple different sources playing different melodies. For this experiment we use as an input a recording of a singing voice accompanied by tubular bells playing a few bars from the round "Frére Jacques". In this case because the spectral characteristics of the two sources are distinct (the harmonic voice vs the inharmonic tubular bells), we need to perform an analysis in which the latent variable assumes two values. This means that we will be estimating two kernel and impulse distributions, each fitting the pattern of each source. The results of the analysis are shown in figure 9. As is evident from the figure the input is a very dense distribution where the included melodies are very hard to spot visually. However the distinct difference between the two spectra representing the two sounds force the two kernel distributions to converge to their shape, and help segment the input in the two instrument parts. Upon examining the impulse distributions we extract we can easily see the structure of the two melodies. Likewise ex-

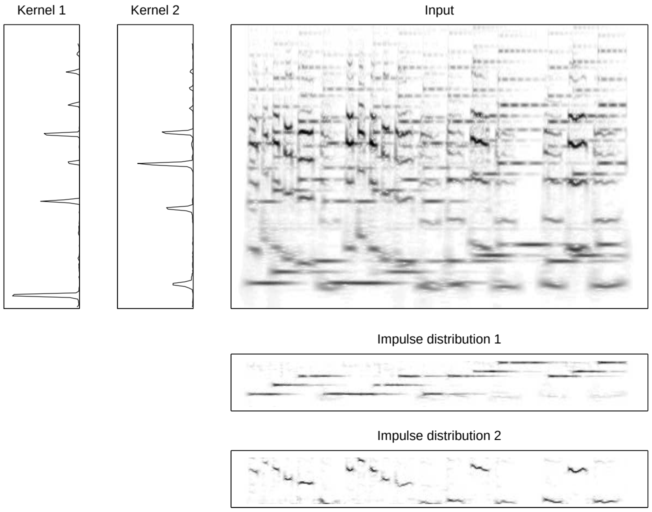


FIG. 9. Analysis of a mixture of two different sources performing simultaneous notes. The analysis results in two kernel and two impulse distributions, each pair describing one of the two sources. The top right plot displays the input. The two left plots show the two extracted kernel distributions, and the two bottom right plots show the impulse distributions that contain the recovered spectral pitch tracks of the two sources. Note that the second and fifth to last notes were performed an octave higher, and the displayed results do not exhibit an octave error.

amining the recovered kernel distributions we can see the two different spectra which represent the characteristics of the two sources in the mixture. The same approach can be applied on mixtures of an arbitrary number of sounds, although as the number of sources increase the disjointedness assumption in the input will gradually be weakened and result in poorer estimates. Having instruments with very similar spectra (e.g. a flute and a clarinet) can also be a problematic case since there will not be sufficient difference between the tracked spectra to easily distinguish them. However using temporal continuity priors or prior knowledge on the structure of the mixture instruments can allow us to offset that problem. This idea is developed in [15]. If the spectra of the mix instruments are sufficiently different it is also possible to use rank estimation methods, such as AIC or BIC, to estimate the number of sources. This is however an intrinsically unreliable approach and is only expected to produce good results in cases with a strong contrast in the character of the instruments in the mix.

The approach of using multiple spectral templates can also be very beneficial when attempting to recover the frequency shifts of a source which is contaminated by additive noise. In this situation we can expect one kernel distribution to latch on to the spectrum of the source we wish to track and another one latching on to the background noise source. The impulse distribution corresponding to the tracked source will again be the spectral pitch track whereas the other impulse distribution will converge to some less structured form that is adequate to describe the presence of noise but will not carry any information about pitch.

If we are provided a way to invert the constant-Q transform, or a similar transform which is also invertible, we can even use this information to selectively reconstruct the input and thus isolate the melody of each source. However, recovering a time waveform from such a decomposition is not straightforward process and we postpone its discussion in future publications.

VI. DISCUSSION

The model we presented is able to overcome some of the challenges we set forth in the introduction of this paper. Although it might seem cumbersome at first, the probabilistic interpretation we have chosen provides a very flexible framework which is easy to extend in multiple ways. In the presence of musical signals one can impose a prior on the structure of the impulse distribution so that it follows the frequency-shift expectations of the musical style at hand. Or if one is interested in temporally smoother spectral pitch tracks, modeling the temporal behavior of a specific kind of source, the application of a dynamical system such as a Kalman filter or a Markov model can incorporate prior temporal knowledge in order to provide more appropriate results [15]. Likewise rules on harmony and counterpoint can enhance this approach to allow polyphonic transcription.

This model is also useful when one knows the spectral characteristics of the sounds that need to be tracked. These characteristics can be applied as a prior on the kernel distribution [15]. Or in the case where these are exactly known, the kernel distributions can be preset and fixed as we only update the impulse distribution. This is essentially a straightforward non-negative deconvolution process. The only complication is that we need to maintain that the output has to be positive and a probability distribution. This results in a more powerful and interpretable representation compared to a cross-correlation output that a straightforward deconvolution would produce.

In conclusion, we presented a frequency-shift tracking model which is flexible enough to deal with situations which can be challenging. This creates a robust front-end for performing pitch tracking which makes soft decisions which can be highly desirable in complex music transcription systems. We presented results which demonstrate the ability of this model to deal with mixtures, inharmonic sounds, and complex tracking situations. We also presented this model in a probabilistic framework which allows clean statistical reasoning and makes it a good candidate for extensions that incorporate statistical priors depending on the input signal.

[1] Kedem, B. 1986. “Spectral analysis and discrimination by zero-crossings,” in proceedings of the IEEE, 74(11):1477–1493, November 1986.

[2] Terez, D. 2002. “Fundamental frequency estimation using signal embedding in state space,” in Journal of the Acoustical Society of America, 112(5):2279.

[3] Moorer, J.A. 1977. “On the transcription of musical sound by computer,” in Computer Music Journal, Vol 1, no. 4, pages 32–38, November 1977.

[4] Brown, J.C. 1992. “Musical fundamental frequency tracking using a pattern recognition method,” in Journal of the Acoustical Society of America, vol. 92, no. 3, pp. 1394–1402, 1992.

[5] Cheveigne, A. and H. Kawahara. 2002. “Yin, a fundamental frequency estimator for speech and music,” in Journal of the Acoustical Society of America, 111(4).

[6] Doval, B. and Rodet, X. 1991. “Estimation of fundamental frequency of musical sound signals,” in International Conference on Acoustics, Speech and Signal Processing, pages 3657–3660.

[7] Goto, M. 2000. “A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings,” in Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, June 2000.

[8] Klapuri, A. P. 1999. “Wide-band pitch estimation for natural sound sources with in-harmonicities,” in Proc. 106th Audio Engineering Society Convention, Munich, Germany, 1999.

[9] Brown, J. C. 1991. “Calculation of a Constant Q Spectral Transform,” in Journal of the Acoustical Society of America vol 89, pages 425-434.

[10] Smaragdis, P., B. Raj, and M.V. Shashanka, 2008. “Sparse and shift-invariant feature extraction from non-negative data,” in proceedings IEEE International Conference on Acoustics and Speech Signal Processing, Las Vegas, Nevada, USA. April 2008

[11] Dempster, A.P, N.M. Laird, D.B. Rubin, 1977. “Maximum likelihood from incomplete data via the EM algorithm” Journal of the Royal Statistical Society, B, 39, 1-38. 1977.

[12] Brand, M.E. 1999. “Structure learning in conditional probability models via an entropic prior and parameter extinction,” Neural Computation, Volume 11 , Issue 5, July 1999.

[13] Robert M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, “On the Lambert W Function, in Advances in Computational Mathematics, volume 5, 1996, pp. 329–359

[14] Shashanka, M.V., B. Raj, and P. Smaragdis. 2007. “Sparse overcomplete latent variable decomposition of counts data,” in Neural Information Processing Systems, December 2007.

[15] Mysore, G. and P. Smaragdis 2009. “Relative pitch estimation of multiple instruments”, *to appear* in proceedings IEEE International Conference on Acoustics and Speech Signal Processing, Taipei, Taiwan. April 2008