CS379H Undergraduate Thesis

Polyphonic Guitar Transcription by Bayesian Methods

Mark Kedzierski
University of Texas at Austin,
Department of Computer Sciences
2005

Table of Contents

Introduction

In this report I will describe the theory, design, and development of a software-based synthesizer driver meant to act the same manner as the Roland GK-2JH guitar synthesizer driver. The GK-2JK is a hardware unit which mounts on the guitar body, below the strings. It picks up the string vibrations, and converts them into messages which can be understood by a synthesizer. The synthesizer then, in turn, plays the messages using the user-selected instrument. This empowers the guitar player to utilize the sounds of thousands of different instruments just by playing guitar. The applications of this capability include music education, composition, and recording. In general, the driver brings the musician and his computer closer together by enabling communication. Any device (software or hardware) which acts in this manner is considered a synthesizer driver.

The role of the driver is to transcribe polyphonic musical melodies. This problem can be divided into instantaneous pitch recognition, note onset/offset detection, and the dispatch of messages which a synthesizer can understand. The detection of pitch in an audio signal is a problem which has long been studied because of its wide range of applications. It is a difficult one however, because the pitch an audio signal is a subjective attribute which cannot be directly measured. Pitch is defined as the characteristic of a sound which places it at a specific location in the musical scale. The ability to detect it varies among different people. Some are born with “perfect pitch”, which means that they can accurately classify the pitch of a sample sound without any reference. Most people however, can only detect intervals, or the difference in pitch between different sounds. Some suggest that pitch detection might be a learned ability, as different cultures often have different musical scales, each with different intervals. In any case, it is difficult to detect any attribute which is, by definition, subjective. The first approach I took to solving this problem was to determine the fundamental frequency of the audio sample, in blocks (windows), and use it to classify its pitch. While this approach resulted in an effective real-time algorithm to detect the pitch of monophonic digital audio, it is ineffective in polyphonic applications.

In this paper I explore a method of transcribing Polyphonic guitar music using probabilistic modeling. First, we define a state space model, based on (Cemgil, 2004), for generating audio signals not unlike those of stringed instruments. We then infer the most likely piano roll which would generate the audio waveform. We do this by generating waveforms for every possible piano roll and comparing them to the source. The closest waveform would give us the correct piano roll. In practice, (because we can't realistically make all those calculations), we do this in real-time as a Viterbi MAP state trajectory estimation problem using Switching Kalman Filters (Murphy, 1998). This means we propagate the most likely sequence of states which led to the current audio sample.

The implementation of my research has been developed in Matlab. The program is capable of learning parameters from sampled audio data, and Viterbi estimation for polyphonic music. I planned to implement the code as a VST plug-in, but didn't have time. VST would allow the product to be used in conjunction with a variety of professional audio software.

The remainder of this document consists of several sections. I have attempted to order the document in such a way that all necessary background theory is presented before its application. However, I found it difficult to express all the background probability while staying on the important issues. In the first section I describe those physical properties of sound waves which are relevant to music. Second, I provide a primer in the music theory necessary to understanding the implementation. Third, I explain the MIDI standard and how it enables digital instruments to communicate. Finally, I outline the solution of a simplified problem, as well as the algorithms/models utilized in the final implementation. To close the document I propose future enhancements the product. I include my bibliography at the end of the document.

Physical Properties of Sound Waves

In the section I discuss the physical nature of sound as relevant in the discussion of musical tones and/or melodies. We define pitch as any the attribute of sound which defines it’s placement on the musical scale. We define a tone as any sound which is associated with a pitch. First I offer a concise explanation by Brian C.J. Moore: “Sound originates from the motion or vibration of an object. This motion is impressed upon the surrounding medium (usually air) as a pattern of changes in pressure. What actually happens is that the atmospheric particles, or molecules, are squeezed closer together than normal (called condensation) and then pulled farther apart than normal (called rarefaction).” (Moore, 1999) We obtain a graphical representation of a sound be plotting the amplitude of pressure variation over time. We call this plot a waveform. I provide an example below. Note because waveforms are a plot against time, they are said to exist in the time-domain.

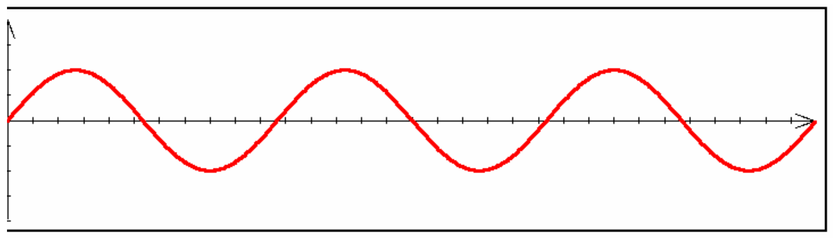


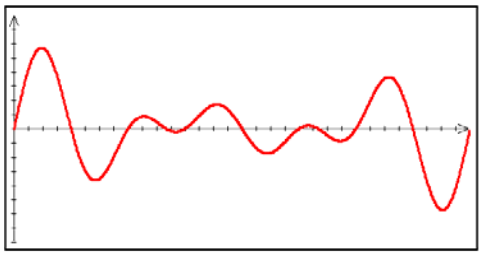
Figure 1: Waveform of a pure tone. Plotted as pressure variation over time.

This waveform is known as a pure-tone. Remember that a tone is defined as any sound which can be classified as having a pitch. A gunshot is not a tone. A pure tone is named for its very “clean” musical sound, very much similar to that of a tuning fork. A pure tone’s waveform is simply a sine wave, whose frequency determines where the note lies on the musical scale. For example, the commonly used A-440 tuning fork produces a wave very similar to a 440 Hz sine wave. We will discuss this in greater detail in the next section.

Figure 1

While pure tones are important in defining pitch and useful in theoretical discussion, we do not encounter them in nature. Instead, we see complex-tones. Like pure-tones, all complex tones are assigned a unique pitch value. But instead of consisting of a single sine wave at a given frequency, they contain several superimposed sine waves, all located at harmonics. More accurately, a complex tone consists of a fundamental frequency and several overtones.

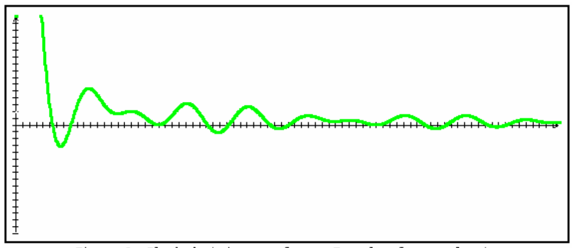
We define the ith harmonic of a given frequency as the ith integral multiple of said frequency. For example, the 0th, 1st, and 2nd harmonics of a 440Hz tone would exist at 440Hz, 880Hz, and 1320Hz, respectively. We define the fundamental frequency of a complex tone to be the 0th harmonic. We similarly define any tones which occur at subsequent harmonics as overtones of that fundamental frequency. The next figure depicts sine waves of frequencies at 3rd, 4th, and 5th harmonics combined to form a complex tone.



Complex tone, containing 3 overtones.

Figure 2

The sound generated from a plucked guitar string is an example of a complex tone. The waveform is shown below. Note how it resembles a damped sin wave, with the amplitude decreasing over time.



Sampled waveform of a plucked guitar string. This is an example of a complex tone.

Figure 3

Music Theory Primer

To begin our primer on music theory, we define a note as any named tone. The ordered collection of all possible notes is known as the musical scale. We define an interval as the spatial distance, on the musical scale, between successive notes. Commonly occurring intervals are given names. The smallest interval is the ½-step (or semitone), which is used to represent the distance between two adjacent notes on the scale. The Whole-step is the equivalent of two ½-steps. Some other commonly occurring intervals are the Fifth, Major Third and Minor Third. We can now define a melody as a series of notes played at varying intervals.

The musical scale is logically divided into groups of twelve notes known as octaves. Every octave consists of twelve semitones which are labeled in the following manner: C, C#/Db, D, D#/Eb, E, F, F#/Gb, G, G#/Ab, A, A#/Bb, B. We use this scheme because each of these labels represents a unique "sound", only twelve of which exists. This means that a C note sounds the "same" as a C note in any other octave. Each octave in the musical scale is labeled by an index, starting at 0. In turn, every note on the musical scale can be uniquely identified by a note label and octave index. Examples are: C0, Db5, and G2. Middle C is represented as C4. So, if we know the pitch of a tone, we know where it belongs on the musical scale. But since real instruments don't make pure tones, we need a way to assign pitch. "Since pitch is a subjective attribute, it cannot be measured directly. Assigning a pitch value to a sound is generally understood to mean specifying the frequency of a pure tone having the same subjective pitch as a sound." (Moore, 1999). We define the relative frequency (ç) of any note ç as the frequency of said pure tone. For example, (C4) = 261.4Hz. Next we discuss how the musical scale maps to the human audible frequency range.

The human audible frequency range is approximately 20-20,000Hz. However, we are not concerned with anything above 5Khz, as we are unable to discern the pitch of frequencies in the that range. (Moore, 1999) The range of 0-5kHz is partitioned into eight octaves, each beginning at C, where (C0)=16.4Hz, (C1)=32.7Hz, (C8)=4186Hz. It may be clear now why C4 is known as middle C. You may notice that (C1) = 2(C0). This leads us to the real definition of the octave as the interval between two tones when their frequencies are in a ratio 2:1 (Moore, 1997) We can deduce from this that the eight octaves are not the same size, instead they increase exponentially. Every interval, in fact, is actually a relationship between two tones. The following table depicts common intervals and their associated ratios:

Interval	**Ratio**
½ step	1:1.05946
Fifth	3:2
Maj Third	5:4
Min Third	6:5

Description of common intervals

Table 1

MIDI Primer

The Musical Instrument Device Interface is widely used by musicians to communicate between digital musical instruments. It is a standard which has been agreed upon by major manufacturers in the industry. It connects not only musical instruments, but all devices such as recorders and lighting systems.

The communication in a MIDI system is done by sending and receiving messages. Examples of some messages related to melody composition are Note-On, Note-Off, Tempo Change, Key-Signature Change, and Time-Signature change. The meanings behind these events are self-explanatory. Each message contains variables specific to the message. For example, a Note-On event would contain information about which note should be played, and the volume at which it should be played. Each note is assigned a specific index. These indices are standard and are listed in Appendix A. The Key-Signature message would contain the new desired key signature.

In practice a synthesizer driver would create and dispatch a series of MIDI messages to a synthesizer. The synthesizer would then playback the desired notes using pre-recorded instrument audio samples.

Bayesian Inference

Bayesian inference problems involve making estimates of the current *state* of a *state-space* system model, given uncertain or incomplete data. This consists of defining a system model and then performing inference algorithms on the model.

■ *State-Space* Models

A *state-space* system model consists of a set of equations which synthesize the output of some physical system. Though not always the case, in this paper I limit discussion to systems which operate/evolve as a function of time (a.k.a. Dynamical Systems).

The system is defined, among other things, by the set of possible 'situations', or *states*, that it can be in at any particular time. The state variable, s_t, represents the current state of the system at each time t, {t: 0≤t≤T}. This can be represented as a single variable, or a vector; with continuous or discrete values. The length of a state vector, $|s_t|$, is labeled as N (N=1 in single variable case). The sequence state variables, $s_{0:T}$, are responsible for generating the output of the system. The actual information stored in the state variable depends on the system being described. In addition to the state variable, all we need to define a system is a *state transition* function, f,

$$\begin{aligned} s_t &= f(s_{t-1}), \\ |s_t| &= N \end{aligned} \tag{1}$$

which describes how the state of the system changes over time. As time progresses through each time-step, the state variable will change to reflect the current time-slice. In practice, the definition of f can depend on a vector of some other constant parameters, θ.

$$s_t = f(s_{t-1}, \theta), \tag{2}$$

Notice that this is a recursive definition. We can then run a simulation of our system over T time-slices, (i.e. calculate the progression of states, $s_{0:T}$), by specifying the base case, s_0, and parameter vector θ.

System Observations

In cases where the *observed* output of a physical system is digitally sampled, the output does not correspond exactly to the set of state vectors. Instead, it is a function of the current state. In this case we add another 'layer' of variables, $y_{1:T}$, to represent the observations. The length of an observation vector, $|y_t|$, is given as M (M=1 in single variable case). Where the value of each observation is given by the *observation function, h.*

$$\begin{aligned} y_t &= h(s_t, \theta) \\ |y_t| &= M \end{aligned} \tag{3}$$

In this case we don't need an initial value, as the definition is not recursive. In summary:

$$\begin{aligned} s_0 &= \pi, \\ s_t &= f(, \theta), \\ y_t &= h(s_t, \theta) \\ |s_t| &= N \\ |y_t| &= M \end{aligned} \tag{4}$$

Linear Dynamical Systems, (LDS's)

If we only consider models with linear transition functions we are dealing with *Linear Dynamical Systems* (LDS's). This linearity enables us to define the transition and observation functions using square transformation matrices A and C, respectively. The new equations are:

$$\begin{aligned} s_t &= A\, s_{t-1} \\ y_t &= C\, s_t, \quad \text{where} \end{aligned} \tag{5}$$

where A is a N×N matrix and C is a M×M matrix. An example of a such a system is a damped sinusoidal with a constant angular frequency ω. The definition of the state-space model follows.

Damped Oscillator Example

The generation of a digitally sampled damped sinusoidal of frequency ω can be stated as a linear dynamical system given by (Cemgil, 2004). We assume that our observations of the system are digital samples taken at a constant rate Fs.

Observations - $y_{1:T}$

At each time slice the observation represents a single audio sample. The value of y_t represents the amplitude of the wave at time t. It has only a single component, so M=1.

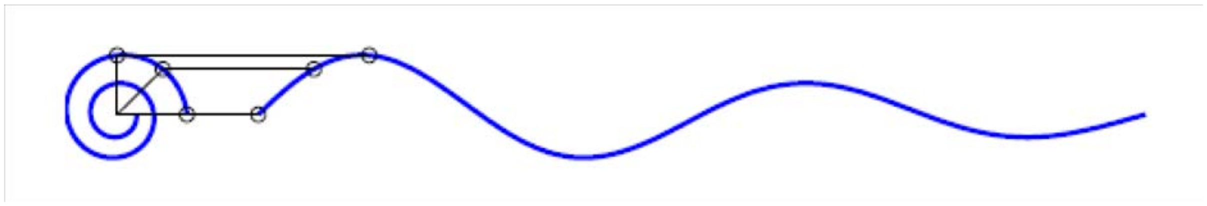
State Vector - $s_{1:T}$

The waves can be generated by a two-dimensional *oscillator* vector, s_t, which rotates around the origin with an angular frequency ω. The length of vector s_t corresponds to the initial amplitude of the sinusoidal and decreases over time at rate constant ρ. This is called the *damping coefficient*. Because the variables ω and ρ stay constant for a given model, they are parameters,

$$\theta = \{\omega, \rho\} \tag{6}$$

System parameters

The observations are generated by projecting the 2-dimensional state vectors onto a 1-dimensional plane (over time). This is depicted in the figure below (Cemgil, 2004).



A damped oscillator in state space form. Left: At each time step, the state vector s_t rotates by ω and its length becomes shorter. Right: The actual waveform is a one dimensional projection from the two dimensional state vector.

Figure 4

Transition Matrix

The transition matrix, A, is responsible for rotating the state vector around the origin by ω and decreasing it's length by ρ (i.e. projecting $s_{t-1} \underset{A}{\Rightarrow} s_t$.). This is defined with Given's rotation matrix, B(ω), which rotates a two dimensional vector by ω radians:

$$B(\omega) \equiv \begin{pmatrix} \text{Cos}[\omega] & -\text{Sin}[\omega] \\ \text{Sin}[\omega] & \text{Cos}[\omega] \end{pmatrix} \tag{7}$$

Given' s rotation matrix.

$$A \equiv \rho B_{\omega}$$

$$A \equiv \begin{pmatrix} \rho\text{Cos}[\omega] & -\rho\text{Sin}[\omega] \\ \rho\text{Sin}[\omega] & \rho\text{Cos}[\omega] \end{pmatrix} \tag{8}$$

State transition matrix

Adding Harmonics

For a system modelling H harmonics, or overtones, we expand the state vector and transition matrices. The length of the state vector will now be 2H. It stores an (x,y) vector for each oscillator in the system; they are appended together.

The oscillator at harmonic h rotates at an angular frequency h ω and is damped with a coefficient ρ^h. The frequencies of overtones are discussed in an earlier section. The decreasing damping factor, (decreasing because $\rho < 1$, and $p_h = (p_{h-1})^2$, along the harmonics is consistent with real-life data from sampled instruments. The new transition matrix is defined as:

$$A \equiv \begin{pmatrix} \rho\text{Cos}[\omega] & -\rho\text{Sin}[\omega] & 0 & 0 & \cdots & \cdots & 0 & 0 \\ \rho\text{Sin}[\omega] & \rho\text{Cos}[\omega] & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \rho^2\,\text{Cos}[2\,\omega] & -\rho^2\,\text{Sin}[2\,\omega] & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \rho^2\,\text{Sin}[2\,\omega] & \rho^2\,\text{Cos}[2\,\omega] & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \rho^H\,\text{Cos}[H\,\omega] & -\rho^H\,\text{Sin}[H\,\omega] \\ 0 & 0 & 0 & 0 & 0 & 0 & \rho^H\,\text{Sin}[H\,\omega] & \rho^H\,\text{Cos}[H\,\omega] \end{pmatrix} \tag{9}$$

Observation Matrix

The observations are simply projections of the state vector. We define the observation matrix as 1×2H dimensional projection matrix:

$$C = [0 \quad 1 \; .... \; 0 \quad 1] \tag{10}$$

Observation matrix

Deterministic Model

Here are our state/observation equations so far:

$$s_t = A\,s_{t-1}$$
$$y_t = C\,s_t$$

Figure 5

Adding some 'randomness'

The above definitions, given s_0 and θ, are enough to describe a fully deterministic model. This means that we can tell *with certainty* what values of $s_{0:T}$, $y_{0:T}$ will be: we can simply plug in the initial variables and calculate each time-slice. While a good start, this is not a realistic model of a sound wave. We need to model both noise in the state transitions, to account for minute frequency/damping variations, as well as observation noise. The latter models the noise introduced by digitizing the wave with recording hardware. This will create a more realistic, i.e. 'noisy', sound wave. We do this adding noise terms to both the state and observation vectors:

$$s_t = A\,s_{t-1} + w \tag{11}$$

$$y_t = C\,s_t + v \tag{12}$$

Modelling process noise

The process noise is defined as *white gaussian*. A noise is white when it is uncorrellated through time; its value at any time t is not dependant to any other time period. It's labeled as *gaussian* because it is drawn from a normal (gaussian) distribution. It particular, a zero-mean gaussian with diagonal covariance matrix Q:

$$w \sim \mathcal{N}\left([0 \quad 0], \begin{pmatrix} Q_{11} & 0 \\ 0 & Q_{22} \end{pmatrix}\right) \tag{13}$$

where $\mathcal{N}(\mu,\Sigma)$ represents a multivariate gaussian distribution with mean μ and covariance matrix Σ.
x ~ $\mathcal{N}(\mu,\Sigma)$ means that the value of x is drawn from the distribution. We can see how the diagonal terms of Q add noise the state vector. We can simplify (10) to get:

$$\begin{aligned} s_t &= A\, s_{t-1} + \mathcal{N}(0, Q) \\ s_t &\sim \mathcal{N}(A\, s_{t-1}, Q) \end{aligned} \tag{14}$$

Modelling observation noise

The definition of the observation noise is identical to the process noise except for the dimensionality. We define a observation noise covariance matrix R and write:

$$\begin{aligned} y_t &= C\, s_t + \mathcal{N}(0, R) \\ y_t &\sim \mathcal{N}(C\, s_t, R) \end{aligned} \tag{15}$$

Modelling state transitions with conditional probability distributions

The noise terms now render the system model indeterministic. This means that we are no longer predict, will full certainty, the values of future state variables. Both the transition and observation functions are continuous variables represented as conditional probability distributions.

$$\Theta = \{\omega, \rho, H\}$$

$$\begin{aligned} s_0 &= \pi \\ s_t &\sim p(s_t \mid s_{t-1}) \equiv \mathcal{N}(A\, s_{t-1}, Q) \\ y_t &\sim p(y_t \mid s_t) \equiv \mathcal{N}(C\, s_t, R) \end{aligned}$$

Where θ represents the given parameter constants, as in (6). π is also given as the initial state.

Table 2

The term $p(y_t \mid s_t)$ is known as the *likelihood* of the observation. It will be explained in later sections.

Hidden / Observed variables

It is important to label the variables in our model as *hidden* or *observed*. The observation terms, $y_{1:T}$, are obviously observed. The state vectors are hidden terms. They are considered *hidden* because we can't directly measure them in any way. We can only derive them past state variables. This is a big problem when we don't have the initial state vector, s_0, from which to calculate s_1..., then s_2; all we have are noisy state observations, $y_{1:T}$. How can we find the values of these hidden state variables?

Estimating values of hidden variables, a.k.a. *filtering*

The answer is to *infer* the values of the hidden variables, $s_{0:T}$, conditioned on noisy observations. This is the most common problem in Bayesian statistics and is known as *filtering*. Think of filtering out the noise and returning the pure source of the observations.

Because the values of the hidden variables are unobserved and indeterministic, we can only make estimates, $\hat{s}_{1:T}$. The hat symbol, ^, indicates an estimated or predicted value.

Because we are making estimates of the hidden variables (a.k.a. *state estimate*, or *belief state*), we also associate a degree of belief with our estimate. This is the essence of Bayesian probability: we use probabilities to represent the certainty in an estimate. More specifically, the *belief state* at time t is a probability density (i.e. normalized distribution) over all possible states at time t given the observations, i.e. $p(s_{1:T} \mid y_{1:T})$. This is known as the *posterior distribution, or filtering density*. It can be calculated using an the prior state estimate, the *likelihood* of the evidence given the estimate, and Bases' Theorem:

$$p(s_{1:T} \mid y_{1:T}) \equiv \frac{p(y_{1:T} | s_{1:T})\, p(s_{1:T})}{p(y_{1:T})} \quad \textbf{(16)}$$

Same equation with labeled terms is:

$$\text{Posterior} \equiv \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \quad \textbf{(17)}$$

The prior probability is propagated from the initial state estimate, $\pi \equiv p(s_0)$,

$$p(s_{1:T}) \equiv \pi\, p(s_1 \mid s_2)\, p(s_2 \mid s_3)\, p(s_3 \mid s_4)\, \ldots\, p(s_{T-1} \mid s_T) \quad \textbf{(18)}$$

The likelihood term is defined in the generative model.

The normalization factor, or *evidence*, is the sum of likelihoods for all states s_t:

$$p(y_{1:T}) \equiv \int_{s_{1:T}} p(y_{1:T} \mid s_{1:T})\, p(s_{1:T}) \quad \textbf{(19)}$$

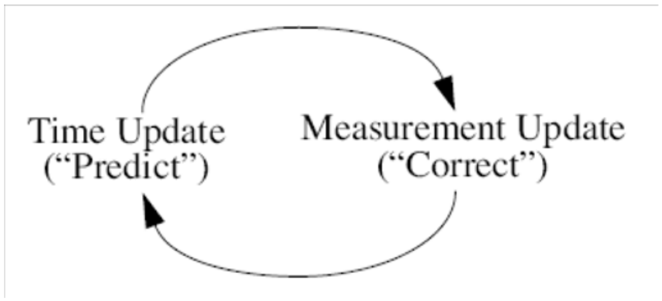
In general, $\sum$ is used to marginalize over discrete variables, $\int$ is used to marginalize continuous variables.

The belief state, $\alpha_{t|t}$, is represented by the first two moments of the posterior state distribution:

$$\begin{aligned} \alpha_{t|t} &\equiv p(s_t \mid y_{1:t}) \\ &\equiv \mathcal{N}(\hat{x}_t, P_t) \equiv \mathcal{N}(E[s_t], E[s_t]\, E[s_t]') \end{aligned} \tag{20}$$

■ Message Propagation

From the recursive definitions above we see that posterior distribution is initially estimated and then propagated across time-slices. This is done in a cycle of prediction and correction equations.



The ongoing discrete Kalman filter cycle. The time update projects the current state estimate ahead in time. The measurement update adjusts the projected estimate by an actual measurement at that time. Figure taken from (Welch, 2001)

Figure 6

Conditioned on the prior distribution from t-1 (or π in case of t=1), a prediction of the next audio sample (a.k.a. *a priori* estimate) can be calculated from the current belief state using the generative model in the predict step:

$$\begin{aligned} \hat{s}_{t|t-1} &= A\, \hat{s}_{t-1} \\ \hat{y}_{t|t-1} &= C\, \hat{s}_{t|t-1} \end{aligned} \tag{21}$$

The projection of the state vector across time slices, $s_{t-1} \xrightarrow{A} s_t$, corresponds to marginalization over s_{t-2}. This is because the state vector at time t only depends s_{t-1}, (i.e. s_t is conditionally independent of $s_{t-2,}$ given s_{t-1}). Updating the estimate covariance, (see Kalman Equation figure below), gives the new *a priori* state estimate,

$$\delta_{t|t-1} \equiv p(s_t \mid s_{t-1},\ y_{1:t-1}), \tag{22}$$

In the correct step, we use the *observation estimate,* $\hat{y}_{t|t-1}$, along with observed sample, y_t, to calculate the conditional likelihood of the observation.

■ What exactly is Likelihood?

The likelihood of a particular audio sample reflects the difference between the predicted audio sample, $\hat{y}_t$, and observed sample, y_t. This difference is known as the *residual*, or the *innovation*:

$$e = y_t - \hat{y}_t, \tag{23}$$

The likelihood is calculated as:

$$\mathcal{L} \quad \equiv \quad \mathcal{N}(e;\, 0,\, V_t) \qquad \textbf{(24)}$$

As the residual increases, the likelihood of the configuration $r_{1:t}$ decreases at a rate inversely proportional to the state estimate covariance, P_t.

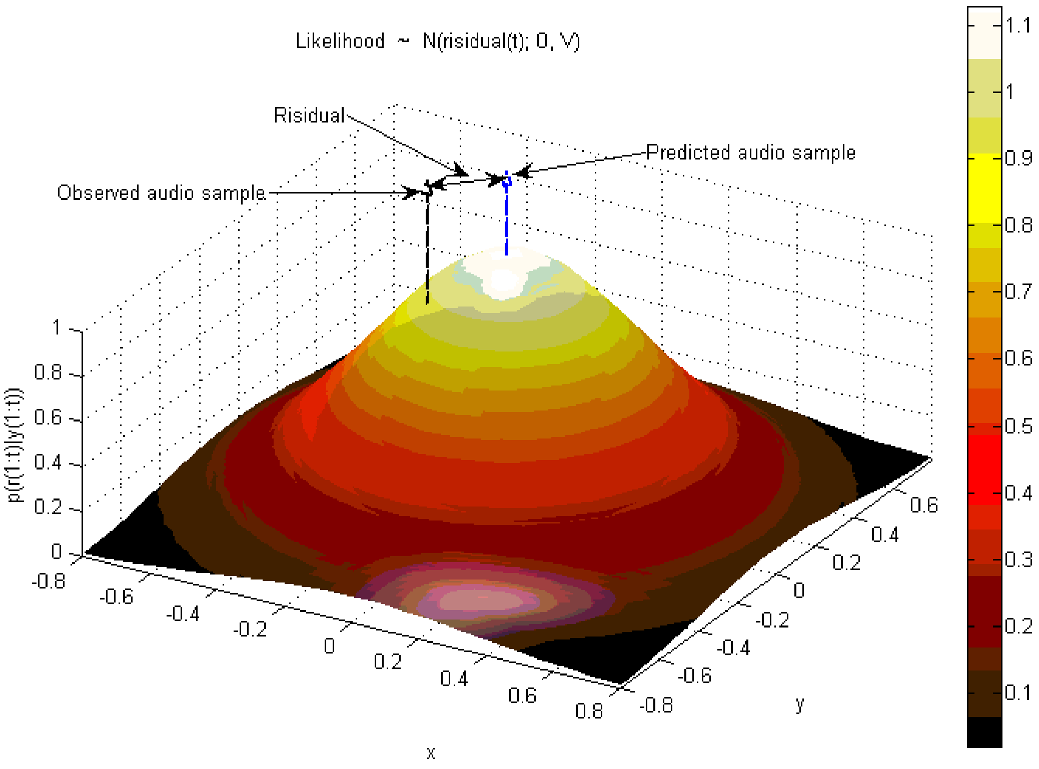


Figure 7

We can update our *a priori* estimate to an *a posteriori* estimate through conditioning on y_t. This corresponds to the *correction* Kalman Filter equations. The correction of the state vector depends on the state estimate covariance.

If the model is properly defined, our estimate covariance will decrease over time (i.e. we can be more confident in our estimate) as our state estimate converges on the source of the noisy samples. If it doesn't, we'll witness observation data with very low likelihoods. At this point we'll want to reconsider our parameter estimates. The full Kalman Filter equations are shown next:

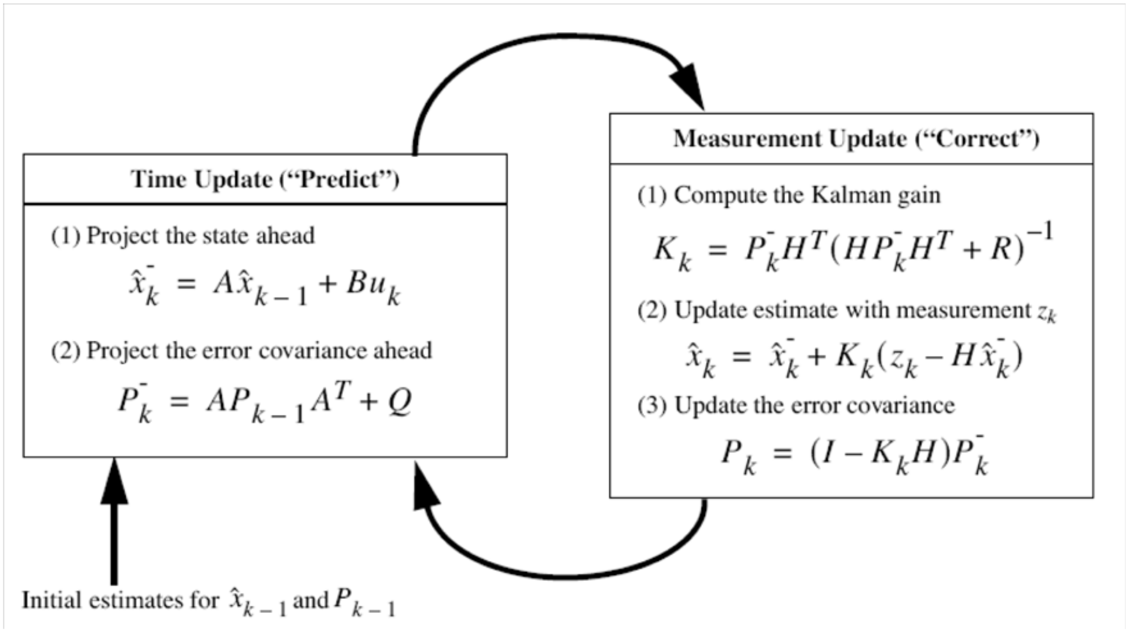

Time Update ("Predict")
(1) Project the state ahead
$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_k$
(2) Project the error covariance ahead
$P_k^- = AP_{k-1}A^T + Q$
Measurement Update ("Correct")
(1) Compute the Kalman gain
$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$
(2) Update estimate with measurement z_k
$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-)$
(3) Update the error covariance
$P_k = (I - K_k H)P_k^-$
Initial estimates for $\hat{x}_{k-1}$ and P_{k-1}

Figure 8

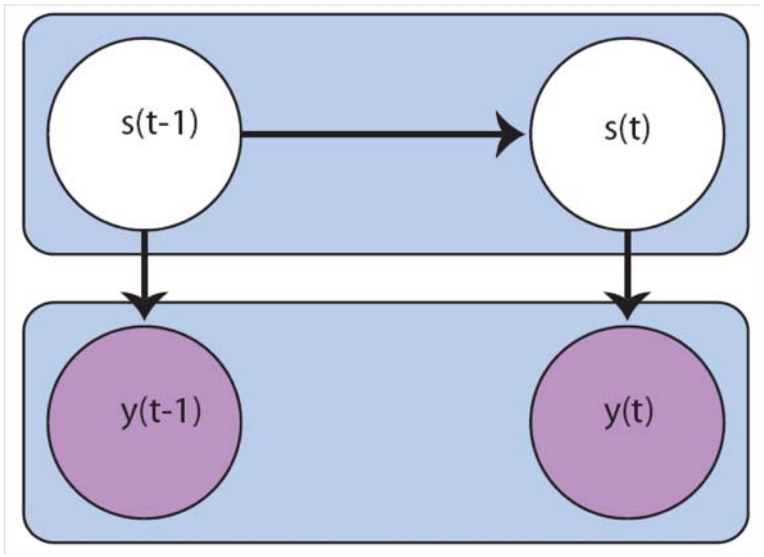
■ Dynamic Bayesian Network, (DBN's)

To show how the prediction and correction steps correspond to marginalization and conditioning, it is helpful to depict the modal graphically as a Dynamic Bayesian Network   A Bayes' net is a directed graph used to represent conditional probability distributions of nodes given their parent nodes. A dynamic Bayesian network (DBN) is a way to extend Bayes nets to model probability distributions over semi-infinite collections of random variables, Z_1, Z_2, ... (Murphy, 2004). We can partition the variables into Zt = (St, Yt) to represent the hidden and output variables of a state-space model (Murphy, 2004).

A DBN is defined to be a pair, $(B_1, B_\rightarrow)$, where B_1 is a BN which defines the prior $p(Z_1)$, and $B_\rightarrow$ is a two-slice temporal Bayes net (2TBN) which defines $p(Z_t \mid Z_{t-1})$ by means of a DAG (directed acyclic graph) as follows (Murphy, 2004):

$$p(Z_t \mid Z_{t-1}) = \prod_{i=1}^{N} p(Z_t^i \mid \mathrm{Pa}(Z_t^i)) \quad (25)$$

We must define the conditional probability distribution (CPD) of each node given its parents. For our model this includes $p(s_0)$, $p(s_t \mid s_{t-1})$ and $p(y_t \mid s_t)$; all are define above. The next figure depicts the DBN for our model:



The top layer of nodes, $s_{1:T}$, represent the state vector of the oscillator generating the note. The bottom layer, $y_{1:T}$, represents the observed audio samples.

Figure 9

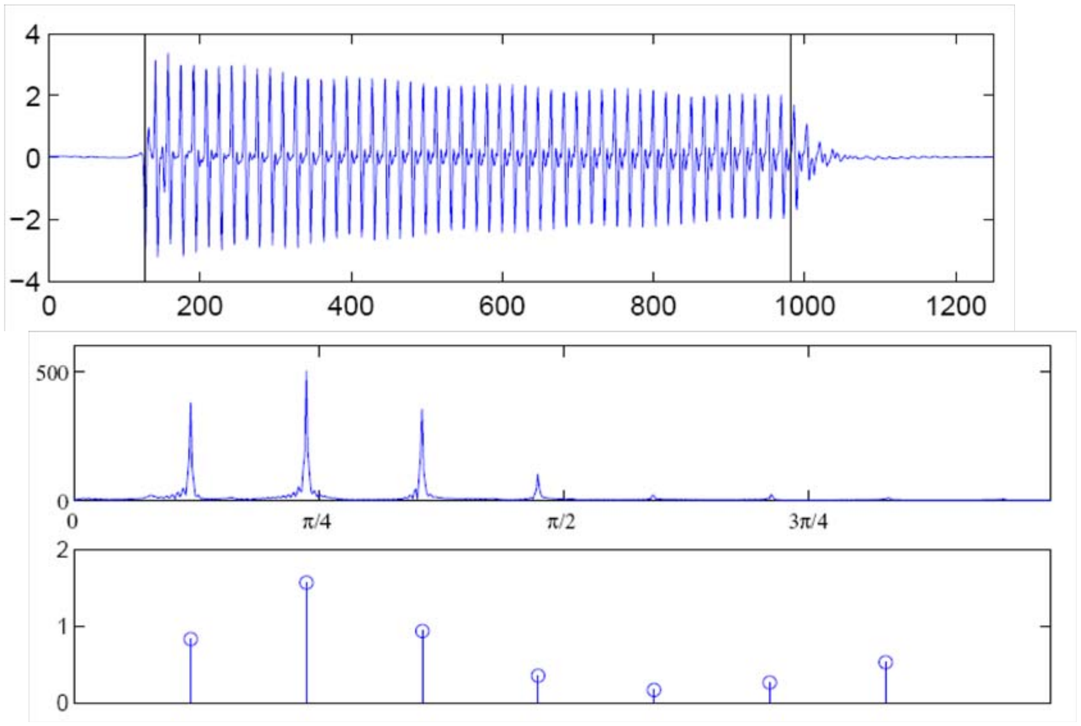
An important point to note is that $p(s_t)$ is conditionally independent of $p(s_{t-2})$ given s_{t-1}. This allows the recursive marginalization needed by tractable filtering.

■ Initial state estimate - π

The initial state estimate, $\pi \equiv p(s_0)$ is drawn from a zero-mean gaussian with covariance matrix S.

$$\pi \sim \mathcal{N}(0, S) \tag{26}$$

The diagonal sum of S, Tr(S), is proportional to the initial amplitude of sound wave. The covariance structure defines how the 'energy' is distributed along the harmonics (Cemgil, 2004). It is initially estimated with a strong 1st and 2nd harmonics. The later overtones have increasingly less energy. This pattern can be shown by a Fourier transform of the audio signal.


Top: Audio signal of electric bass. Originally at 22 kHz, downsampled by factor D=20. Middle: Frequency Spectrum (DFT) of audio signal
Bottom: Diagonal entries of S for H=7 correspond to overtone amplitudes

■ Switching Kalman Filter Model for monophonic melodies

In this section the model is expanded to model a piano roll representation. The *piano roll*, or *score*, is a sequence indicators which label each time-slice with a state of *sound*, or *mute*. We define values *sound*=1 and *mute*=2 for simplicity.

■ Transition Matrices

We need to define transition matrices, $\{A_{\text{sound}}, A_{\text{mute}}\}$, for each state of the oscillator. They are the same as before except for the damping coefficient:

$$\begin{aligned} A_{\text{sound}} &\equiv \rho_{\text{sound}}\, B\,(\omega) \\ A_{\text{mute}} &\equiv \rho_{\text{mute}}\, B\,(\omega) \end{aligned} \tag{27}$$

ρ_{sound} is slightly less than 1. ρ_{mute} is much smaller than ρ_{sound}.

■ Defining the switch variables - $r_{1:T}$

The variables are referred to as *switch* variables because, at each time-step, they select which of two transition functions is used to project the next state. The two transition functions represent the *sound* and *mute* states:

$$r_{1:T} \quad \equiv \{ \quad r_t : \quad r_t \in \{\text{sound} \wedge \text{mute}\}, \quad 0 \le t \le T\} \tag{28}$$

For every 2-slice period, there are four possible combinations of $r_{t-1:t}$. The distribution of the configurations defines the prior on piano rolls. They can be organized as:

$$p(r_1) \quad \equiv \quad u(\{\text{sound, mute}\}) \quad \equiv \quad [.5,\ .5] \tag{29}$$

$$\begin{aligned} p\,(i,\ j) &\equiv p\,(r_t = i \mid r_{t-1} = j) \\ &\equiv \begin{Bmatrix} p(r_t = 1 \mid r_{t-1} = 1) & p(r_t = 1 \mid r_{t-1} = 2) \\ p(r_t = 2 \mid r_{t-1} = 1) & p(r_t = 2 \mid r_{t-1} = 2) \end{Bmatrix} \\ &\equiv \begin{pmatrix} 1 - 10^{-7} & 10^{-7} \\ 10^{-7} & 1 - 10^{-7} \end{pmatrix} \end{aligned} \tag{30}$$

The initial prior is a uniform distribution. The conditional distribution is a multinomial where it is much more likely for the switch variable to remain the same over time slices. We also define the function, isonset_t. It is true if there has been a note onset at the current time slice. A note onset is represented by the mute to sound transition. The updated model is as follows:

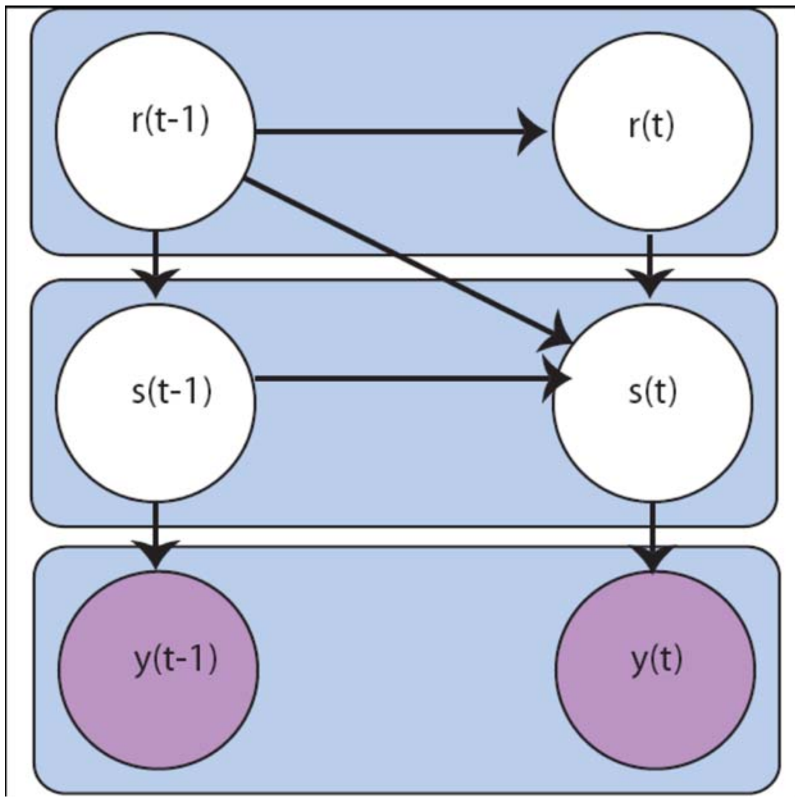
$$\theta = \{\omega,\ \rho_{\text{sound}},\ \rho_{\text{mute}},\ H\}$$

$$\text{isonset}_t \ \equiv \ [r_t = \text{sound} \ \wedge \ r_{t-1} = \text{mute}]$$

$$\begin{aligned} r_1 &\equiv u(\{\text{sound, mute}\}) \equiv \{.5,\ .5\} \\ r_t &\equiv p(i,\ j) \\ s_0 &\sim \mathcal{N}(0,\ S) \\ s_t &\sim \mathcal{N}(A_{r_t}\, s_{t-1},\ Q) \\ y_t &\sim \mathcal{N}(C\, s_t,\ R) \end{aligned}$$

Table 3

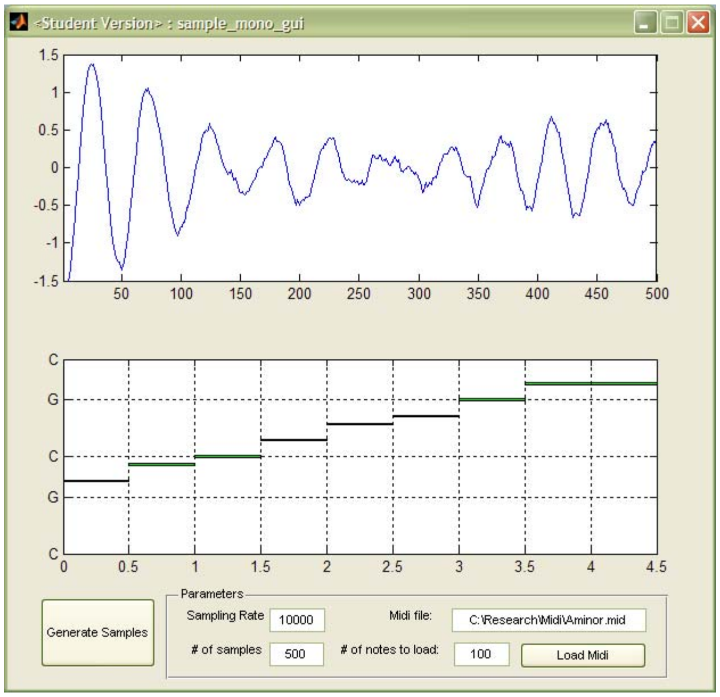
Below is the DBN:



The top layer of nodes, $s_{1:T}$, represent the state vector of the oscillator generating the note. The bottom layer, $y_{1:T}$, represents the observed audio samples.

■ Model sample generation in Matlab

The following figure depicts samples taken from the model using Matlab.



Parameters

Sampling Rate 10000

Midi file: C:\Research\Midi\Aminor.mid

# of samples 500

# of notes to load: 100

Figure 12

■ Inference for the Switching Kalman Filter model

■ Viterbi Estimation

Given an audio waveform, our goal is to infer the most likely sequence of piano roll indicators, $r^*_{1:T}$, which could give rise to the observed audio samples. This is, in general, a Viterbi estimation problem. The value we are seeking is defined as the *Maximum A Posteriori* trajectory:

$$r^*_{1:T} \equiv \underset{r_{1:T}}{\operatorname{argmax}}\, p(r_{1:T} \mid y_{1:T}) \tag{31}$$

It represents the most likely sequence of hidden variables to cause an observed output. This is different from the posterior distribution because it is just a point estimate. Calculation is the same as for filtering, except for replacing the summation over r_{t-2} with the maximization (MAP). It is important to note that, in Viterbi estimation, we propagate a *filtering potential*, $\delta_{1:T}$, as opposed to a *filtering density*, $\alpha_{1:T}$. The term potential used to indicate that this value is not normalized. Because we only care about the best piano roll, (i.e. configuration $r_{t:T}$ with highest likelihood), we can save calculations by using a point estimate.

$$p(r_{1:T} \mid y_{1:T}) \propto \underbrace{\int_{s_t} p(y_{1:t} \mid s_t, r_{1:t})\, p(s_{1:t} \mid r_{1:t})\, p(r_{1:t})}_{\text{filtering potential}} \tag{32}$$

Message Propagation

Inference is more difficult in the switching state-space model because we have to keep track of every possible enumeration of $r_{1:T}$ and return the sequence with the highest likelihood. The representing of the filtering potential is a Mixture of Gaussians (MoG) with a single component for each enumeration of $r_{1:T}$, (we use H=1 to simplify the examples):

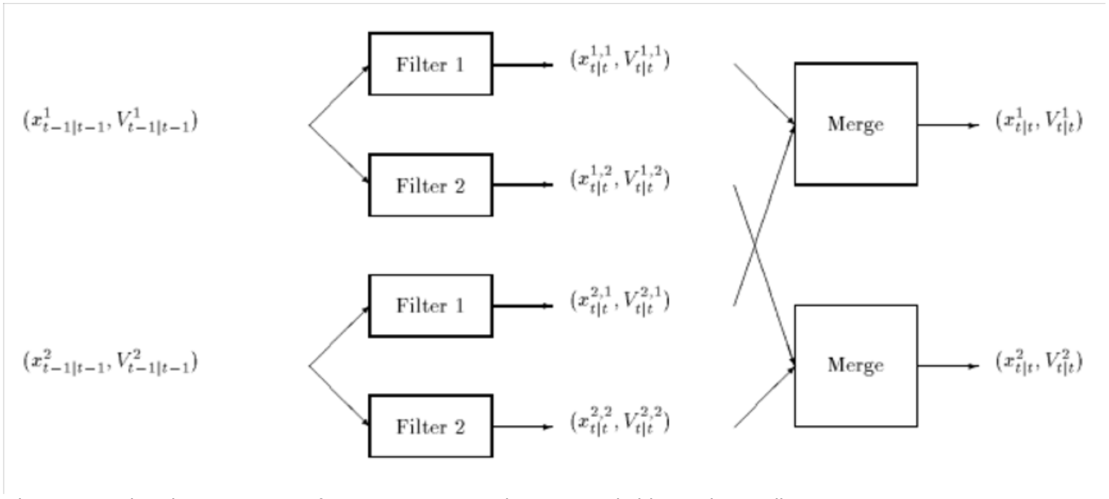
$$\delta_{t|t} \equiv p(y_{1:t}, s_t, r_t, r_{t-1}) \equiv \left\{ \begin{matrix} \delta_t(1, 1) & \delta_t(1, 2) \\ \delta_t(2, 1) & \delta_t(2, 2) \end{matrix} \right\} \tag{33}$$

where each,

$$\begin{aligned} \delta_{t|t}(i, j) &\equiv p(y_{1:t}, s_t, r_t = i, r_{t-1} = j) \\ &\equiv p(y_{1:t}, s_t, r_t = i, r_{t-1} = j) \end{aligned} \tag{34}$$

is also a MoG.

At each time step we make separate estimates for every configuration of piano roll indicators $\{r_t, r_{t-1}\}$. This corresponds to the prediction step. It can also be considered an *expand* step. As shown by the left side of the next figure.


$(x^1_{t-1|t-1}, V^1_{t-1|t-1})$
$(x^2_{t-1|t-1}, V^2_{t-1|t-1})$
Filter 1
Filter 2
Filter 1
Filter 2
$(x^{1,1}_{t|t}, V^{1,1}_{t|t})$
$(x^{1,2}_{t|t}, V^{1,2}_{t|t})$
$(x^{2,1}_{t|t}, V^{2,1}_{t|t})$
$(x^{2,2}_{t|t}, V^{2,2}_{t|t})$
Merge
Merge
$(x^1_{t|t}, V^1_{t|t})$
$(x^2_{t|t}, V^2_{t|t})$

Show Expand and Merge steps of message propagation on a Switching Kalman Filter

Figure 13

The correction/marginalization over r_{t-2} is done in the *merge* step. In the case of filtering, this is done with summation. In case of Viterbi, we use maximization.

Prediction

Upon receiving $\delta_{t-1|t-1}$ from the previous time slice,

$$\delta_{t-1|t-1} \equiv \left\{ \begin{matrix} \delta_{t-1|t-1}(1, 1) & \delta_{t-1|t-1}(1, 2) \\ \delta_{t-1|t-1}(2, 1) & \delta_{t-1|t-1}(2, 2) \end{matrix} \right\} \equiv p(y_{1:t-1}, s_{t-1}, r_{t-1}, r_{t-2}) \tag{35}$$

we with to calculate the *a priori* estimate, $\delta_{t|t-1}$, as follows.

First we marginalize over r_{t-2}. This is done by union over each row of $\delta_{t-1|t-1}$. These terms represent

$$\xi^i_{t-1} \equiv p(y_{1:t-1}, s_{t-1}, r_{t-1} = i) \equiv \sum_{r_{t-2}} p(y_{1:t-1}, s_{t-1}, r_{t-1} = i, r_{t-2}) \quad (36)$$

This marginalization corresponds to the 'merge' box in the above figure.

For components $\delta(1,1)$, $\delta(1,2)$, and $\delta(2,2)$ we next apply the transition model, $p(s_t \mid y_{1:t-1}, s_{t-1}, r_{t-1}, r_{t-2})$, using the Kalman Filter Time Update equations. We label each predicted potential as $\psi_{i,j}$.

$$\begin{gathered}\psi_{i,j} \equiv \int_{s_t-1} p(s_t \mid y_{1:t-1}, s_{t-1}, r_{t-1})\, \xi^j_{t-1}, \\ r_t = i\end{gathered} \quad (37)$$

Multiplying each potential by the prior, $p(r_t \mid r_{t-1})$, gives:

$$p(y_{1:t-1}, s_t, r_t, r_{t-1}) \equiv p(y_{1:t-1}, s_t, r_{t-1}) p(r_t \mid r_{t-1}) \quad (38)$$

The component, $\delta_t(1, 2)$, is different because it represents a note onset. It's important to note that the piano roll configuration before the onset, $r_{1:\text{onset}}$, doesn't effect the likelihood of future indicators after it. This enables us to we can replace messages from ξ^{mute}_{t-1} with the maximum (scalar) likelihood estimate among them. We introduce this scalar value, Z^{mute}_{t-1}, as a prior for the next onset and tag the message with $r^*_{1:t-1}$. This replacement renders the algorithm tractable. The *a priori* potential is be given as:

$$\delta_{t|t-1} \equiv \left\{ \begin{matrix} p(1, 1)\, \psi_{1,1} & p(1, 2)\, Z^{\text{mute}}_{t-1}\, \xi^j_{t-1} \\ p(2, 1)\, \psi_{2,1} & p(2, 2)\, \psi_{2,2} \end{matrix} \right\} \quad (39)$$

■ Correction

Finally, the a Posteriori state estimate, $\delta_{t|t}$, is calculated using the Kalman Filter Correction equations on each component of the a *priori estimate:*

$$\delta_{t|t} \equiv p(y_{1:t}, s_t, r_t, r_{t-1}) \equiv p(y_t \mid s_t) p(y_{1:t-1}, s_t, r_t, r_{t-1}) \quad (40)$$

. The next section shows the algorithm running in Matlab.

■ Forward Message Representation

To propagate the MAP trajectory, we define our forward-filtering message, δ_t, to store other variables in addition to the state estimate.

--Potential---

$\hat{s}_t$	$\equiv \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$, state estimate	
P_t	$\equiv$ 2x2 estimate cov.	
$P_{t\|t-1}$	$\equiv$ 2-slice state covariance matrix	
p_t	$\equiv p(r_{1:t})$	$\equiv$ prior probability
K_t	$\equiv$ Kalman Filter Gain	
V_t	$\equiv$ innovation covariance	

$y_t - \hat{y}_t \equiv$ residual

$\mathcal{L} \equiv p(y_{1:t} \mid r_{1:t}) = \mathcal{N}(y_t; \hat{y}_t, V_t) \quad \equiv$ *likelihood*

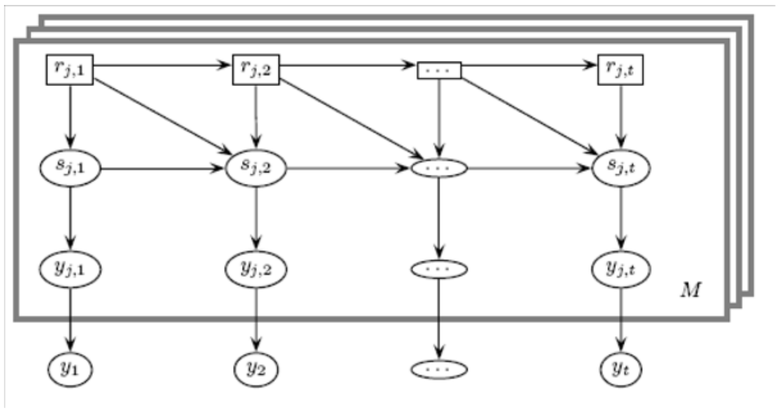
$p(r_{1:t} \mid y_{1:t}) \equiv$ posterior

potential

$$\propto \int_{s_t} p(y_{1:t} \mid s_t, r_{1:t})\, p(s_{1:t} \mid r_{1:t})\, p(r_{1:t})$$

■ Extending the Model for multiple notes - $j_{1:T}$

We can simply expand the model to work for several notes at different frequencies. The frequencies are given, and are set to common MIDI values. We modify piano roll indicators $r_{1:T}$ to be $r_{1:T,1:M}$, where M is the number of notes in our model. Inference is the same except for indexing (Cemgil, 2004). The new DBN is as follows:

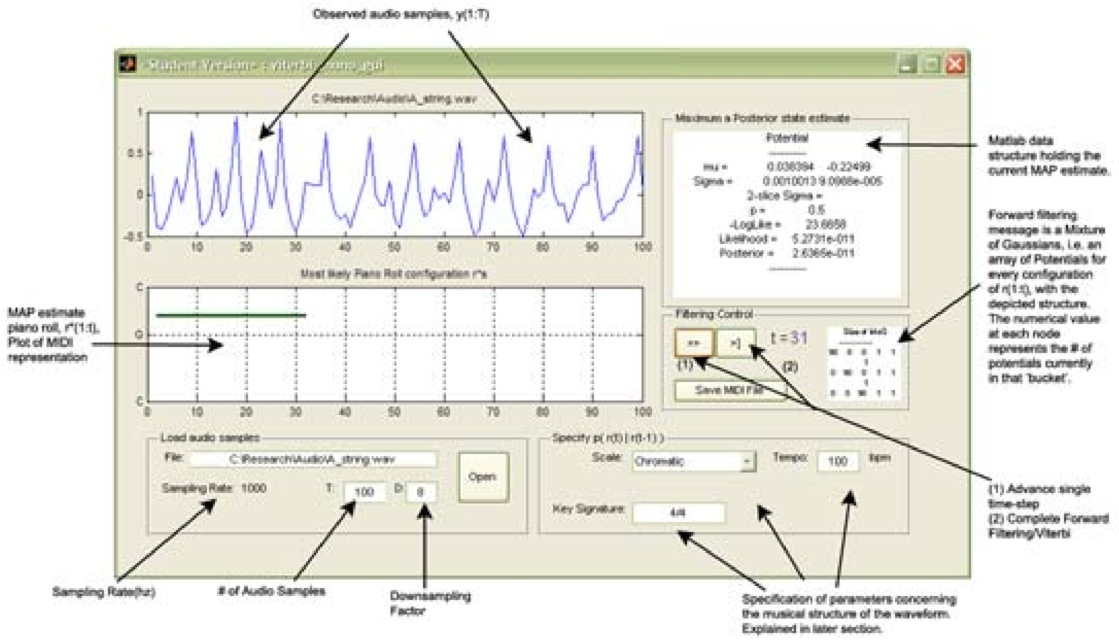


Graphical Model. The rectangle box denotes "plates", M replications of the nodes inside. Each plate, j = 1, . . . ,M represents the sound generator (note) variables through time. (Cemgil, 2004)

Figure 14

■ Matlab Implementation

Below in a labeled screenshot of my Matlab implementation of Viterbi estimation.



Observed audio samples, y(1:T)

Matlab data structure holding the current MAP estimate.

Forward filtering message is a Mixture of Gaussians, i.e. an array of Potentials for every configuration of r(1:t), with the depicted structure. The numerical value at each node represents the # of potentials currently in that 'bucket'.

MAP estimate piano roll, r*(1:t). Plot of MIDI representation

(1) Advance single time-step
(2) Complete Forward Filtering/Viterbi

Specification of parameters concerning the musical structure of the waveform. Explained in later section.

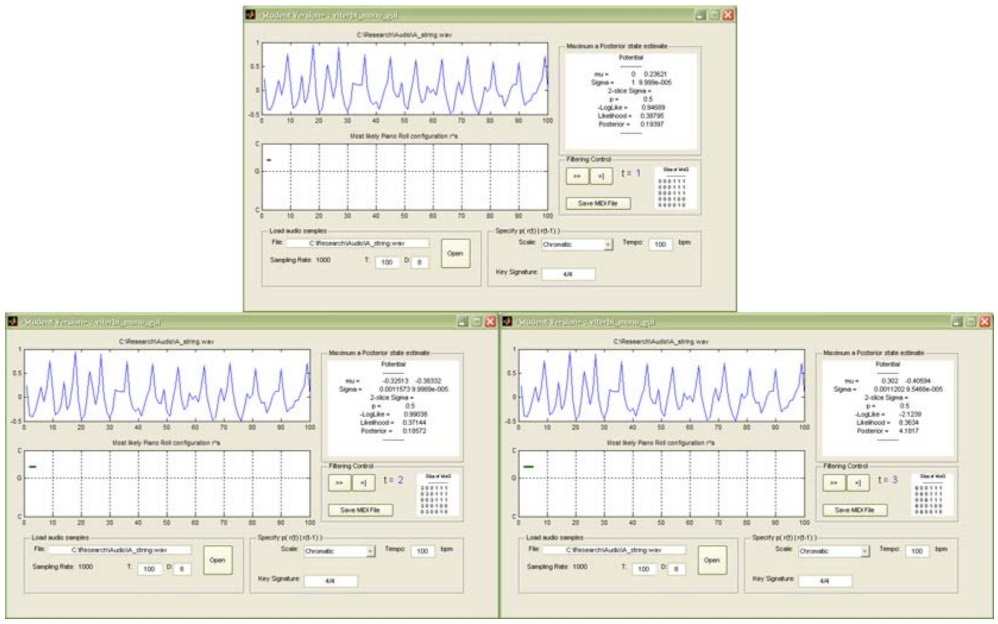
```
Maximum a Posterior state estimate
            Potential
            ---------
mu =          0.036394    -0.22499
Sigma =       0.0010013 9.0988e-005
         2-slice Sigma =
p =                0.5
-LogLike =         23.6658
Likelihood =    5.2731e-011
Posterior =     2.6365e-011
            ---------
```

Load audio samples
File: C:\Research\Audio\A_string.wav
Sampling Rate: 1000 T: 100 D: 8

Specify p(r(t) | r(t-1))
Scale: Chromatic Tempo: 100 bpm
Key Signature: 4/4

Figure 15

Below are screenshots from the first three time slices of Viterbi calculation:



```
t = 1
mu =          0    0.23621
Sigma =       1 9.999e-005
         2-slice Sigma =
p =                0.5
-LogLike =      0.94689
Likelihood =    0.38795
Posterior =     0.19397
```

```
t = 2
mu =          -0.32513    -0.38332
Sigma =       0.0011573 9.9989e-005
         2-slice Sigma =
p =                0.5
-LogLike =      0.99038
Likelihood =    0.37144
Posterior =     0.18572
```

```
t = 3
mu =          0.302    -0.40594
Sigma =       0.0011202 9.5468e-005
         2-slice Sigma =
p =                0.5
-LogLike =      -2.1239
Likelihood =    8.3634
Posterior =     4.1817
```

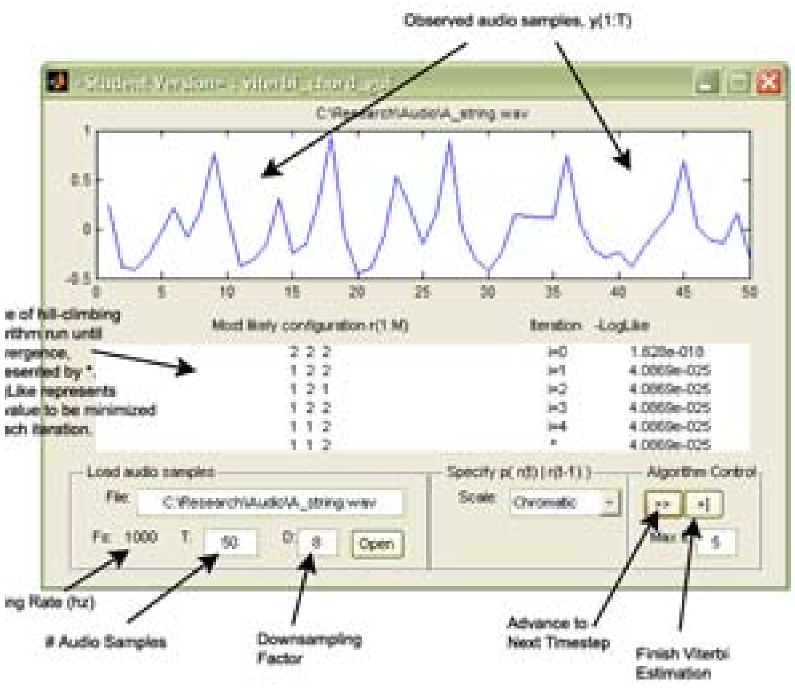
Figure 16

Polyphonic Chord Transcription

In this section we use the Viterbi estimation techniques learned in the last section use them to transcribe polyphonic chords. The difference in this case, is that we infer the variables $r_{1:M}$, rather than $r_{1:M,1:T}$. We assume that the configuration stays constant for t=0:T. This means that we have to work in windows if we wish to detect changes in chords over time.

The algorithm, given in (Cemgil, 2004), is a simple greedy search. We start with at an initial estimate of the chord configuration (zero's or drawn from prior), and prior $p(r_{1:M})$. A more informative prior can be very helpful in finding the correct configuration, $r_{1:M}$. We then calculate the likelihood of our estimate, α, and all configurations differing from our by a single note. If α is the most likely, we stop (at the local maximum). Otherwise we continue until convergence.

Using a uniform prior, $p(r_{1:M})$. We are (usually) able to find the correct chord configuration by taking the best of three trials using different initial estimates. This productivity can be improved by making more specifications in the prior, as explained in the next section. This is done by specifying some properties of the composition which generated the waveform. Next I show my Matlab implementation:



```
Most likely configuration r(1:M)        Iteration  -LogLike
            2  2  2                        i=0     1.620e-018
            1  2  2                        i=1     4.0869e-025
            1  2  1                        i=2     4.0869e-025
            1  2  2                        i=3     4.0869e-025
            1  1  2                        i=4     4.0869e-025
            1  1  2                        *       4.0869e-025
```

File: C:\Research\Audio\A_string.wav
Fs: 1000 T: 50 D: 8
Scale: Chromatic
Max: 5

The Matlab GUI for polyphonic chord transcription. The various controls and values are labeled.

Figure 17

Extending the Prior Distribution

The uniform prior over piano rolls/chord configurations leaves room for improvement. We can improve performance by specifying more information about the creation of the audio waveform. Because were are considering the use of this implementation mainly as a musical composition tool we can assume that the performer of the audio samples is available to supply the new parameters, which represent the details of the underlying composition.

■ Tonality in the prior

The parameters concerning the composition include the diatonic scale, Φ, the tempo, Δ, and the time-signature, Ξ. The diatonic scale labels each note, j_t, as a member or non-member. Non-members are relatively very unlikely, especially as the note duration increases. This is true because the use of these notes is usually limited to *passing* (transition) tones; they are rarely held ringing for a long duration. They are also more common on *offbeats* rather than *onbeats* (see Δ term). Members are additionally weighted by their index in the scale. For example, the I, IV, and V notes usually appear more commonly than others. In general, the probability of a note at time t is given by,

$$\mathrm{p}(\ j_t \mid j_{t-1}, \Phi, \Xi, \Delta\ ) \tag{41}$$

The tempo term, Δ, labels each time-slice as *onbeats* or *offbeats,* given a specified quantization factor. In general, note onsets are much more common when they occur on or near a beat onset. The time-signature, Ξ, expands on this by considering exactly where a given note appears in the composition.

■ Physical instrument properties

We can also use physical properties of the guitar used to create the audio samples. This is possible because there are only a limited number of distinct fretboard *fingerings* for a given note/chord. For example, most notes appear at only 1-3 locations on the guitar fretboard, and each string can only generate a single simultaneous note.

For monophonic melodies, we define a fretboard-distance heuristic which weights the likelihood of a particular note by it's distance, in frets, from the previous note; the closer notes being more common. We also define notes appearing on adjacent strings as more likely.

For chords we use a similar heuristic by adding a weight to each configuration, proportional to the number of distinct hand *fingerings*.

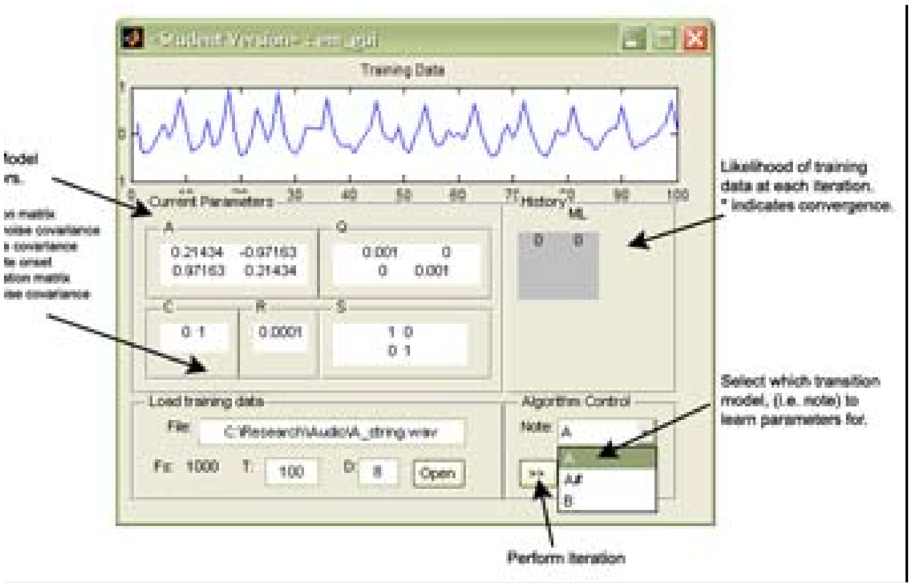
Instrument Parameter Learning

To achieve reliable performance in real applications, our estimates of system model parameters, θ, can be learned for a particular instrument/recording setup. This is done by applying the standard EM (Expectation Maximization) algorithm to training data for each of M notes sampled from each string of the guitar.

The EM algorithm iteratively maximizes the likelihood of a particular training set by calculating the derivative of the likelihood term, for each parameter, and adjusting appropriately. The likelihood of the training data is guaranteed to increase at every iteration until convergence. For a more complete description see (Murphy, 1998), (Digalakis, 1993), (Ghahramani, Z, 1996), and (Cemgil, 2004).

We create the training data as follows. For each string, we sample the instrument playing each note at a regular interval and set tempo. For example, we could use a tempo of 100 bpm and play 12 ascending notes landing on beats. This would mean playing the first 12 frets on the guitar in ascending order. We then use an algorithm to split up each string waveform into sections, $y_{1:T}^{1:J}$, each containing a single note.

Each section is created such that the transition model remains constant. (i.e. note onset occurs at t=1 with j remaining constant, $r_{1:T,j}$=1). This permits us to treat each section as a *non-switching* linear dynamical system and calculate the likelihood using standard Kalman Filter equations. My Matlab implementation, with important values labels, is shown below.



A: 0.21434 -0.97163 / 0.97163 0.21434
Q: 0.001 0 / 0 0.001
C: 0 1
R: 0.0001
S: 1 0 / 0 1
File: C:\Research\Audio\A_string.wav
Fs: 1000 T: 100 D: 8
Note: A (A, A#, B)

The Matlab GUI for EM parameter learning. The various controls and values are labeled.

Figure 18

Bibliography

[1] Ahrendt, Peter. *The Multivariate Gaussian Probability Distribution*. IMM, Technical University of Denmark

[2] Barber, D. (2004). *A Stable Switching Kalman Smoother*. In IDIAP-RR 04-89.

[3] Cemgil, A. T., Kappen, H. J., & Barber, D. (2003). *Generative model based polyphonic music transcription*. In *Proc. of IEEE WASPAA*, New Paltz, NY. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

[4] Cemgil, A. T. *Bayesian Music Transcription*. PhD thesis, Radboud University of Nijmegen, 2004.

[5] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. *ML estimation of a stochastic linear systemswith the EM algorithm and its application to speech recognition*. *IEEE Trans. on Speech and Audio Proc.*, 1(4):421-442, 1993.

[6] Doucet, A., de Freitas, N., K. Murphy, and S. Russell. *Rao-blackwellised particle filtering for dynamic Bayesian networks*. In *UAI*, 2000.

[7] Fearnhead, P. (2003). *Exact and efficient bayesian inference for multiple changepoint problems.* Tech. rep., Dept. of Math. and Stat., Lancaster University.

[8] Ghahramani, Z., & Hinton, G. E. (1996). *Parameter estimation for linear dynamical systems*. (crg-tr-96-2). Tech. rep., University of Totronto. Dept. of Computer Science.

[9] Kalman, R. E. (1960). *A new approach to linear filtering and prediction problems*. Transaction of the ASME-Journal of Basic Engineering, 35-45.

[10] Kedzierski, Mark (2004). *Real-Time Monophonic Melody Transcription*. Undergraduate research rep, University of Texas at Austin, Deparment of Computer Science.

[11] MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

[12] Moore, Brian C.J. (1997). *An Introduction to the Psychology of Hearing 4th Edition*, (Academic Press, London).

[13] Murphy, K. P. (1998). *Switching Kalman filters*. Tech. rep., Dept. of Computer Science, University of California, Berkeley.

[14] Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California, Berkeley.

[15] Raphael, C., & Stoddard, J. (2003). *Harmonic analysis with probabilistic graphical models*. In *Proc. ISMIR*.

[16] Welch, G., & Bishop, G. (2001). *An Introduction to the Kalman Filter*. SIGGRAPH course 8, University of North Carolina at Chapel Hill.