Rhythm Quantization for Transcription

Ali Taylan Cemgil*; Peter Desain†; Bert Kappen*
*SNN, University of Nijmegen, The Netherlands
†NICI, University of Nijmegen, The Netherlands
cemgil@mbfys.kun.nl

August 4, 1999

Abstract

Automatic Music Transcription is the extraction of an acceptable notation from performed music. One important task in this problem is rhythm quantization which refers to categorization of note durations. Although quantization of a pure mechanical performance is rather straightforward, the task becomes increasingly difficult in presence of musical expression, i.e. systematic variations in timing of notes and in tempo. For transcription of natural performances, we employ a framework based on Bayesian statistics. Expressive deviations are modelled by a probabilistic performance model from which the corresponding optimal quantizer is derived by Bayes theorem. We demonstrate that many different quantization schemata can be derived in this framework by proposing suitable prior and likelihood distributions. The derived quantizer operates on short groups of onsets and is thus flexible both in capturing the structure of timing deviations and in controlling the complexity of resulting notations. The model is trained on data resulting from a psychoacoustical experiment and thus can mimic the behaviour of a human transcriber on this task.

1 Introduction

Automatic Music Transcription is the extraction of an acceptable musical description from performed music. The interest into this problem is motivated by the desire to design a program, which creates automatically a notation from a performance. In general, e.g. when directly operating on an acoustical recording of polyphonic music (polyphonic pitch tracking), this task proved to be a very difficult one and stays yet as an unsolved engineering problem. Surprisingly, even a virtually simpler subtask still remains difficult, namely, producing an acceptable notation from a list of onset times (e.g. a sequence of MIDI events) under unconstrained performance conditions.

Although quantization of a "mechanical" performance is rather straightforward, the task becomes increasingly difficult in presence of expressive variations, which can be thought as systematic deviations from a pure mechanical performance. In such unconstrained performance conditions, mainly two types of systematic deviations from exact values do occur. At small time scale notes can be played accented or delayed. At large scale tempo can vary, for example the musician(s) can accelerate (or decelerate) during performance or slow down (ritard) at the end of the piece. In any case, these timing variations usually obey a certain structure since they are mostly intended by the performer. Moreover, they are linked to several attributes of the performance such as meter, phrase, form, style etc. (Clarke, 1985). To devise a general computational model (i.e. a performance model) which takes all these factors into account, seems to be quite hard.

Another observation important for quantization is that we perceive a rhythmic pattern not as a sequence of isolated onsets but rather as a perceptual entity made of onsets. This also suggests that attributes of neighboring onsets such as duration, timing deviation etc. are correlated in some way.

This correlation structure is not fully exploited in commercial music packages, which do automated music transcription and score type setting. The usual approach taken is to assume a constant tempo throughout the piece, and to quantize each onset to the nearest grid point implied by the tempo and a suitable pre-specified minimum note duration (e.g. eight, sixteenth etc.). Such a grid quantization schema implies that each onset is quantized to the nearest grid point *independent* of its neighbours and thus all of its attributes are assumed to be independent, hence the correlation structure is not employed. The consequence of this restriction is that users are required to play along with a fixed metronome and without any expression. The quality of the resulting quantization is only satisfactory if the music is performed according to the assumptions made by the quantization algorithm. In the case of grid-quantization this is a mechanical performance with small and independent random deviations.

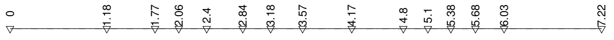
More elaborate models for rhythm quantization indirectly take the correlation structure of expressive deviations into account. In one of the first attempt to quantization, Longuet-Higgins (1987) described a method in which he uses hierarchical structure of musical rhythms to do quantization. Desain et al. (1992) use a relaxation network in which pairs of time intervals are attracted to simple integer ratios. Pressing and Lawrence (1993) use several template grids and compare both onsets and inter-onset intervals (IOI's) to the grid and select the best quantization according to some distance criterion. The Kant system Agon et al. (1994) developed at IRCAM uses more sophisticated heuristics but is in principle similar to (Pressing and Lawrence, 1993).

The common critic to all of these models is that the assumptions about the expressive deviations are implicit and are usually hidden in the model, thus it is not always clear how a particular design choice effects the overall performance for a full range of musical styles. Moreover it is not directly possible to use experimental data to tune model parameters to enhance the quantization performance.

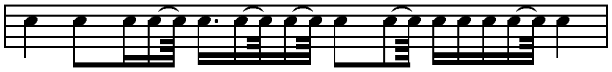
In this paper we describe a method for quantization of onset sequences. The paper is organized as follows: First, we state the transcription problem and define the terminology. Using the Bayesian framework we briefly introduce, we describe probabilistic models for expressive deviation and notation complexity and show how different quantizers can be derived from them. Consequently, we train the resulting model on experimental data obtained from a psychoacoustical experiment and compare its performance to simple quantization strategies.

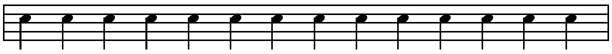
2 Problem Description

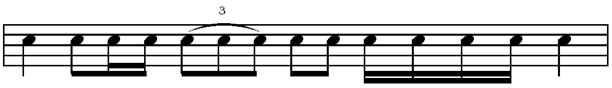
We defined automated music transcription as the extraction of an *acceptable* description (music notation) from a music performance. In this study we concentrate on a simplified problem, where we assume that a list of onset times is provided excluding tempo, pitch or note duration information. Given any sequence of onset times, we can in principle easily find a notation (i.e. a sequence of rational numbers) to describe the timing information arbitrarily well. Equivalently, we can find several scores describing the same rhythmic figure for any given error rate, where by error we mean some distance between onset times of the performed rhythm and the mechanical performance (e.g. as would be played by a computer). Consider the performed simple rhythm in Figure 1(a) (from Desain and Honing (1991)). A very fine grid quantizer produces a result similar to Figure 1(b). Although this is a very accurate representation, the resulting notation is far too complex. Another extreme case is the notation in Figure 1(c). Although this notation is simple, it is very unlikely that it is the intended score, since this would imply unrealistic tempo changes during the performance. Musicians would probably agree that the "smoother" score shown in Figure 1(d) is a better representation. This example suggests that a *good score* must



(a) Example: A performed onset sequence



(b) "Too" accurate quantization. Altough the resulting notation represents the performance well, it is unacceptably complicated.

(c) "Too" simple notation. This notation is simpler but is a very poor description of the rhythm.



(d) Desired quantization balances accuracy and simplicity.

Figure 1: Different Quantizations of an onset sequence.

be "easy" to read while representing the timing information accurately. This is apparently a trade-off and a quantization schema must balance these two conflicting requirements. In the following section we will more concretely define what we mean by a simple score and accurate representation.

3 Rhythm Quantization Problem

3.1 Definitions

In this section we will give formal definitions of the terms that we will use in the derivations to follow. A *performed rhythm* is denoted by a sequence $[t_i]$[1] where each entry is the time of occurrence of an onset. For example, the performed rhythm in Figure 1(a) is represented by $t_1 = 0$, $t_2 = 1.18$, $t_3 = 1.77$, $t_4 = 2.06$ etc. We will also use the terms *performance* or *rhythm* interchangeably when we refer to an onset sequence.

A very important subtask in transcription is tempo tracking, i.e. the induction of a sequence of points (i.e. *beats*) in time, which coincides with the human sense of rhythm (e.g. foot tapping) when listening to music. Significant research has already been done on psychological and computational modeling aspects of this behavior (Large, 1995; Toiviainen, 1999).

We call such a sequence of beats a *tempo track* and denote it by $\vec{\tau} = [\tau_j]$ where τ_j is the time at which j'th beat occurs. We note that for automatic transcription, $\vec{\tau}$ is to be estimated from $[t_i]$.

Once a tempo track $\vec{\tau}$ is given, the rhythm can be segmented into a sequence of segments, each of duration $\tau_j - \tau_{j-1}$. The onsets in the j'th segment are normalized and denoted by $\mathbf{t}_j = [t_j^k]$ for all $\tau_{j-1} \leq t_i < \tau_j$ where

$$t_j^k = \frac{t_i - \tau_{j-1}}{\tau_j - \tau_{j-1}} \tag{1}$$

Here $k = 1 \dots K_j$ where K_j denotes the number of onsets in the j'th segment[2]. In other words the onsets are scaled and translated such that an onset just at the end of the segment is mapped to one and another just at the beginning to zero. The segmentation of a performance is given in Figure 2.

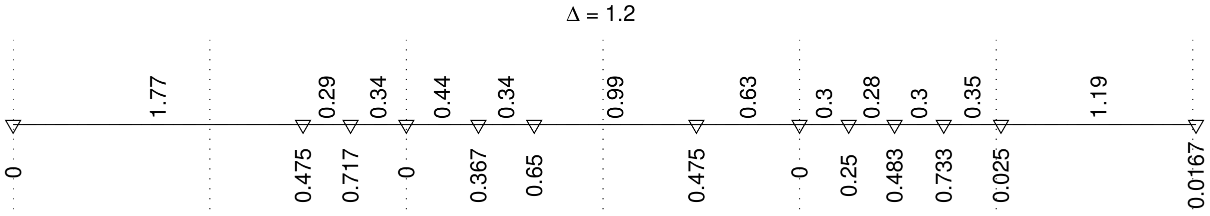


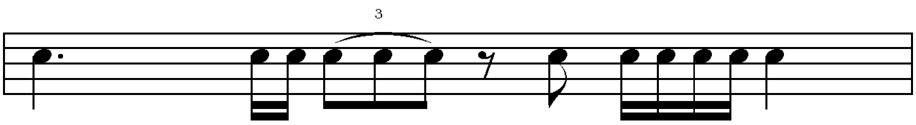
Figure 2: Segmentation of a performance by a tempo track (vertical dashed lines) $\vec{\tau} = [0.0, 1.2, 2.4, 3.6, 4.8, 6.0, 7.2, 8.4]$. The resulting segments are $\mathbf{t}_0 = [0]$, $\mathbf{t}_1 = [0.475, 0.717]$ etc.

Once a segmentation is given, quantization reduces to mapping onsets to locations, which can be described by simple rational numbers. Since in western music tradition, notations are generated by recursive subdivisions of a whole note, it is also convenient to generate possible onset quantization locations by regular subdivisions. We let $\mathcal{S} = [s_i]$ denote a subdivision schema, where $[s_i]$ is a sequence of small prime numbers. Possible quantization locations are generated by subdividing the unit interval $[0, 1]$. At each new iteration i, the intervals already generated are divided further into s_i equal parts and the resulting endpoints are added to a set C. Note that this procedure places the quantization locations on a grid of points c_n where two

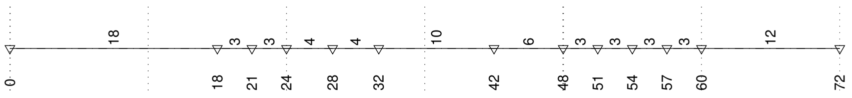
---

[1] We will denote a set with the typical element x_j as $\{x_j\}$. If the elements are ordered (e.g. to form a string) we will use $[x_j]$.

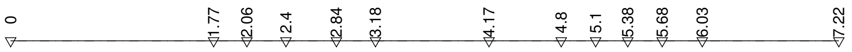
[2] When an argument applies to all segments, we will drop the index j.



(a) Notation



(b) Score



(c) Performance

Figure 3: A simplified schema of onset quantization. A notation (a) defines a score (b) which places onsets on simple rational points with respect to a tempo track (vertical dashed lines). The performer "maps" (b) to a performance (c). This process is not deterministic; in every new performance of this score a (slightly) different performance would result. A performance model is a description of this stochastic process. The task of the transcriber is to recover both the tempo track and the onset locations in (b) given (c).

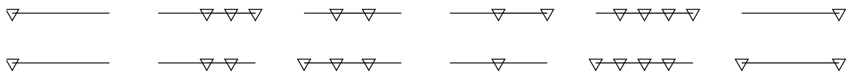


Figure 4: Two equivalent representations of the notation in Figure 3(a) by a code vector sequence

neighboring grid points have the distance $1/\prod_i s_i$. We will denote the first iteration number at which the grid point c is added to C as the *depth* of c with respect to $\mathcal{S}$. This number will be denoted as $d(c|\mathcal{S})$.

As an example consider the subdivision $\mathcal{S} = [3, 2, 2]$. The unit interval is divided first into three equal pieces, then the resulting intervals into 2 and etc. At each iteration, generated endpoints are added to the list. In the first iteration, 0, 1/3, 2/3 and 1 are added to the list. In the second iteration, 1/6, 3/6 and 5/6 are added, etc. The resulting grid points (filled circles) are depicted in Figure 5. The vertical axis corresponds to $d(c|\mathcal{S})$.

If a segment $\mathbf{t}$ is quantized (with respect to $\mathcal{S}$), the result is a K dimensional vector with all entries on some grid points. Such a vector we call a *code vector* and denote as $\mathbf{c} = [c_k]$, i.e. $\mathbf{c} \in C \times C \cdots \times C = C^K$. We call a set of code-vectors a *codebook*. Since all entries of a code vector coincide with some grid points, we can define the *depth of a code vector* as

$$d(\mathbf{c}|\mathcal{S}) = \sum_{c_k \in \mathbf{c}} d(c_k|\mathcal{S}) \tag{2}$$

A score can be viewed as a *concatenation* of code vectors $\mathbf{c}_j$. For example, the notation in Figure 3(a) can be represented by a code vector sequence as in Figure 4. Note that the representation is not unique, both code vector sequences represent the same notation.

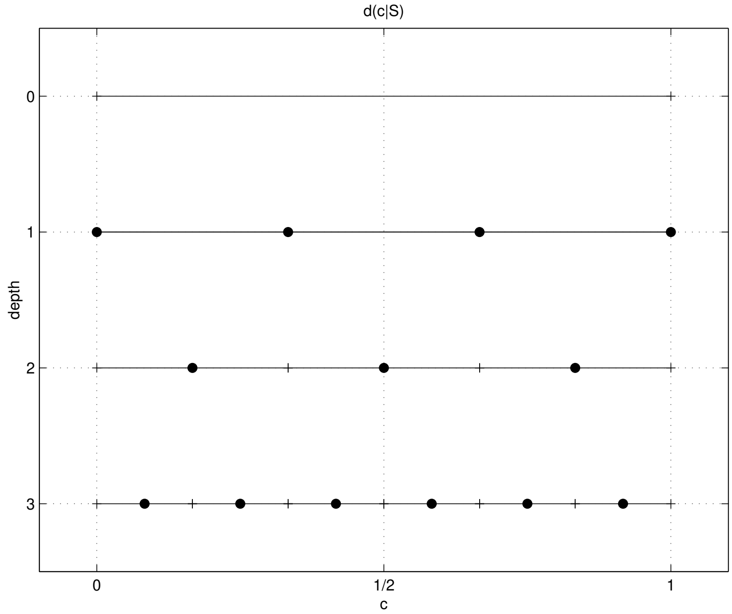


Figure 5: Depth of gridpoint c by subdivision schema $\mathcal{S} = [3, 2, 2]$

3.2 Performance Model

As described in the introduction section, natural music performance is subject to several systematic deviations. In lack of such deviations, every score would have only one possible interpretation. Clearly, two natural performances of a piece of music are never the same, even performance of very short rhythms show deviations from a strict mechanical performance. In general terms, a *performance model* is a mathematical description of such deviations, i.e. it describes how likely it is that a score is mapped into a performance (Figure 3). Before we describe a probabilistic performance model, we briefly review a basic theorem of probability theory.

3.3 Bayes Theorem

The joint probability $p(A, B)$ of two random variables A and B defined over the respective state spaces S_A and S_B can be factorized in two ways:

$$p(A, B) = p(B|A)p(A) = p(A|B)p(B) \tag{3}$$

where $p(A|B)$ denotes the conditional probability of A given B: for each value of B, this is a probability distribution over A. Therefore $\sum_A p(A|B) = 1$ for any fixed B. The marginal distribution of a variable can be found from the joint distribution by summing over all states of the other variable, e.g.:

$$p(A) = \sum_{B \in S_B} p(A, B) = \sum_{B \in S_B} p(A|B)p(B) \tag{4}$$

It is understood that summation is to be replaced by integration if the state space is continuous. Bayes theorem results from Eq. 3 and Eq. 4 as:

$$p(B|A) = \frac{p(A|B)p(B)}{\sum_{B \in S_B} p(A|B)p(B)} \tag{5}$$

This rather simple looking "formula" has surprisingly far reaching consequences and can be directly applied to quantization. Consider the case that B is a score and S_B is the set of all possible scores. Let A be the observed performance. Then Eq 5 can be written as

$$\begin{aligned} p(\text{Score}|\text{Performance}) &\propto p(\text{Performance}|\text{Score}) \times p(\text{Score}) && (6) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} && (7) \end{aligned}$$

which combines the goodness of fit of the performance to the score (the likelihood) with the prior belief in the score to give the posterior belief in the model after we see the data. In this framework, a performance model is the conditional probability distribution $p(\text{Performance}|\text{Score})$. Finding the best score given some performance requires the simultaneous optimization of the performance model and the prior.

By using our definitions of a code vector $\mathbf{c}$ and a segment $\mathbf{t}$, Eq. 6 is equivalent to a maximum a-posteriori (MAP) estimation problem given as

$$p(\mathbf{c}|\mathbf{t}) \quad \propto \quad p(\mathbf{t}|\mathbf{c})p(\mathbf{c}) \tag{8}$$

where the best code vector $\mathbf{c}^*$ is given by

$$\mathbf{c}^* \quad = \quad \underset{\mathbf{c} \in \mathbf{C}^{\mathbf{K}}}{\operatorname{argmax}}\, p(\mathbf{c}|\mathbf{t}) \tag{9}$$

We can also define a related quantity $\mathcal{L}$ (minus log-posterior) and try to minimize this quantity rather then maximizing Eq. 8 directly. This simplifies the form of the objective function without changing the locations of local extrema since $\log(x)$ is a monotonically increasing function.

$$\mathcal{L} \quad = \quad -\log p(\mathbf{c}|\mathbf{t}) \propto -\log p(\mathbf{t}|\mathbf{c}) + \log \frac{1}{p(\mathbf{c})} \tag{10}$$

The $-\log p(\mathbf{t}|\mathbf{c})$ term in Equation 10 can be interpreted as a distance measuring how far the rhythm is played from the perfect mechanical performance. The $\log \frac{1}{p(\mathbf{c})}$ term, which is large when the prior probability $p(\mathbf{c})$ of the codevector is low, can be interpreted as a complexity term, which penalizes complex notations. The best quantization balances these two terms in an optimal way.

The form of a performance model can be in general very complicated. However, in this article we will consider a subclass of performance models where the expressive timing is assumed to be an additive noise component which depends on $\mathbf{c}$. The model is given by

$$\mathbf{t}_j = \mathbf{c}_j + \varepsilon_j \tag{11}$$

where ε_j is a vector which denotes the *expressive timing deviation*. In this paper we will assume that ε is normal distributed with zero mean and covariance matrix $\Sigma_\varepsilon(\mathbf{c})$, i.e. the correlation structure depends upon the code vector. We denote this distribution as $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon(\mathbf{c}))$. Note that when ε is the zero vector, ($\Sigma_\varepsilon \to \mathbf{0}$), the model reduces to a so-called "mechanical" performance.

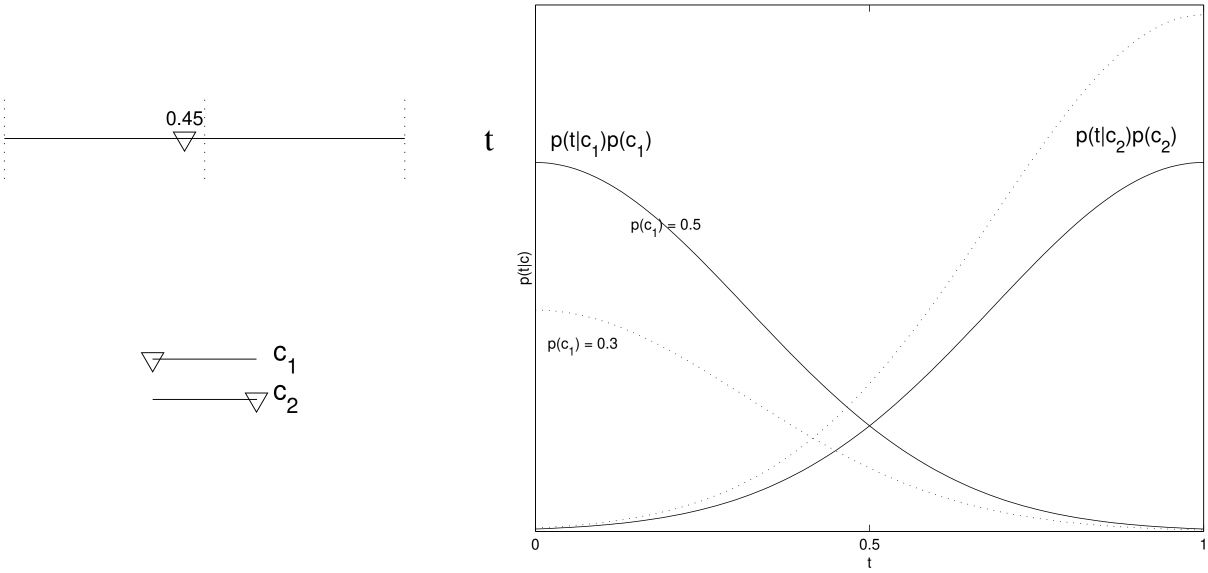


Figure 6: Quantization of an onset as Bayesian Inference. When $p(c) = [1/2, 1/2]$, at each t, the posterior $p(c|t)$ is proportional to the solid lines, and the decision boundary is at $t = 0.5$. When the prior is changed to $p(\mathbf{c}) = [0.3, 0.7]$ (dashed), the decision boundary moves towards 0.

3.4 Example 1: Scalar Quantizer (Grid Quantizer)

We will now demonstrate on a simple example how these ideas are applied to quantization.

Consider a one-onset segment $\mathbf{t} = [0.45]$. Suppose we wish to quantize the onset to one of the endpoints, i.e. we are using effectively the codebook $\mathbf{C} = \{[0], [1]\}$. The obvious strategy is to quantize the onset to the nearest grid point (e.g. a grid quantizer) and so the code-vector $\mathbf{c} = [0]$ is chosen as the winner.

The Bayesian interpretation of this decision can be demonstrated by computing the corresponding likelihood $p(\mathbf{t}|\mathbf{c})$ and the prior $p(\mathbf{c})$. It is reasonable to assume that the probability of observing a performance $\mathbf{t}$ given a particular $\mathbf{c}$ decreases with the distance $|\mathbf{c} - \mathbf{t}|$. A probability distribution having this property is the normal distribution. Since there is only one onset, the dimension $K = 1$ and the likelihood is given by

$$p(t|c) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(t-c)^2}{2\sigma^2})$$

If both codevectors are equally probable, a flat prior can be choosen, i.e. $p(c) = [1/2, 1/2]$. The resulting posterior $p(c|t)$ is plotted in 6. The decision boundary is at $t = 0.5$, where $p(c_1|t) = p(c_2|t)$. The winner is given as in Eq. 9

$$c^* = \operatorname*{argmax}_c p(c|t)$$

Different quantization strategies can be implemented by changing the prior. For example if $\mathbf{c} = [0]$ is assumed to be less probable, we can choose another prior, e.g. $p(\mathbf{c}) = [0.3, 0.7]$. In this case the decision boundary shifts from 0.5 towards 0 as expected.

3.5 Example 2: Vector Quantizer

Assigning different prior probabilities to notations is only one way of implementing different quantization strategies. Further decision regions can be implemented by varying the conditional probability distribution $p(\mathbf{t}\,|\mathbf{c})$. In this section we will demonstrate the flexibility of this approach for quantization of groups of onsets.

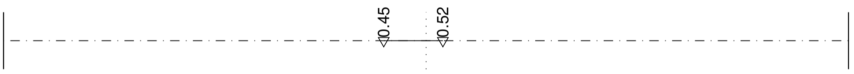


Figure 7: Two Onsets

Consider the segment $\mathbf{t} = [0.45, 0.52]$ depicted in Figure 7. Suppose we wish to quantize the onsets again only to one of the endpoints, i.e. we are using effectively the codebook $\mathbf{C} = \{[0,0],[0,1],[1,1]\}$. The simplest strategy is to quantize every onset to the nearest grid point (e.g. a grid quantizer) and so the code-vector $\mathbf{c} = [0,1]$ is the winner. However, this result might be not very desirable, since the inter-onset interval (IOI) has increased more than 14 times, (from 0.07 to 1). It is less likely that a human transcriber would make this choice since it is perceptually not very realistic. We could try to solve this problem by employing another strategy : If $\delta = t_2 - t_1 > 0.5$, we use the code-vector $[0,1]$. If $\delta \leq 0.5$, we quantize to one of the code-vectors $[0,0]$ or $[1,1]$ depending upon the average of the onsets. In this strategy the quantization of $[0.45, 0.52]$ is $[0,0]$.

Although considered to be different in the literature, both strategies are just special cases which can be derived from Eq. 10 by making specific choices about the correlation structure (covariance matrix Σ_ε) of expressive deviations. The first strategy assumes that the expressive deviations of both onsets are independent of each other. This is apparantly not a very realistic model for timing deviations in music. The latter corresponds to the case where onsets are linearly dependent; it was assumed that $t_2 = t_1 + \delta$ and only δ and t_1 were considered in quantization. This latter operation is merely a linear transformation of onset times and is implied by the implicit assumption about the correlation structure. Indeed some quantization models in the literature focus directly on IOI's rather then on onset times.

More general strategies, which can be quite difficult to state verbally, can be specified by different choices of Σ_ε and $p(\mathbf{c})$. Some examples for the choice $\Sigma_\varepsilon = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and constant $p(\mathbf{c})$ are depicted in Figure 8. The ellipses denote the set of points which are equidistant from the center and the covariance matrix Σ_ε determines their orientation. The lines denote the decision boundaries. The interested reader is referred to Duda and Hart (1973) for a discussion of the underlying theory.

3.5.1 Likelihood for the Vector Quantizer

For modeling the expressive timing ε in a segment containing K onsets, we propose the following parametric form for the covariance matrix

$$\Sigma_\varepsilon(\mathbf{c}) = \sigma^2 \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,K} \\ \rho_{1,2} & 1 & \rho_{n,m} & \vdots \\ \vdots & \rho_{n,m} & \ddots & \vdots \\ \rho_{1,K} & \cdots & \cdots & 1 \end{pmatrix} \tag{12}$$

where

$$\rho_{n,m} = \eta \exp(-\frac{\lambda^2}{2}(c_m - c_n)^2) \tag{13}$$

Here, c_m and c_n are two distinct entries (grid points) of the code vector $\mathbf{c}$. In Eq. 13, η is a parameter between -1 and 1, which adjust the amount of correlation strength between two onsets. The other parameter λ adjusts the correlation as a function of the distance between entries in the code vector. When λ is zero, all entries are correlated by the equal amount, namely η. When λ is large, the correlation approaches rapidly to zero with increasing distance.

This particular choice for $p(\varepsilon)$ reflects the observation that onsets, which are close to each other, tend to be highly correlated. This can be interpreted as follows: if the onsets are close to each other, it is easier to quantify the IOI and then select an appropriate translation for the onsets by keeping the IOI constant. If the grid points are far away from each other, the correlation tends to be weak (or sometimes negative), which suggests that onsets are quantized independently of each other. In section 4, we will verify this choice empirically.

3.5.2 Prior for the Vector Quantizer

The choice of the prior $p(\mathbf{c})$ reflects the complexity of codevector $\mathbf{c}$. In this article we propose a complexity measure from a probabilistic point of view. In this measure, the complexity of a codevector $\mathbf{c} = [c_i]$ is determined by the depth of c_i with respect to the beat (See Eq. 2) and the time signature of the piece. See Figure 9.

The prior probability of a code-vector with respect to $\mathcal{S}$ is chosen as

$$p(\mathbf{c}|\mathcal{S}) \propto e^{-\gamma d(\mathbf{c}|\mathcal{S})} \tag{14}$$

Note that if $\gamma = 0$, then the depth of the codevector has no influence upon its complexity. If it is large, (e.g. $\gamma \approx 1$) only very simple rhythms get reasonable probability mass. This choice is also in accordance with the intuition and experimental evidence: simpler rhythms are more frequently used then complex ones. The marginal prior of a codevector is found by summing out all possible subdivision schemes.

$$p(\mathbf{c}) = \sum_{\mathcal{S}} p(\mathbf{c}|\mathcal{S})p(\mathcal{S}) \tag{15}$$

where $p(\mathcal{S})$ is the prior distribution of subdivision schemas. For example, one can select possible subdivision schemas as $\mathcal{S}_1 = [2,2,2]$, $\mathcal{S}_2 = [3,2,2]$, $\mathcal{S}_3 = [2,3,2]$. If we have a preference towards the time signature (4/4), the prior can be taken as $p(\mathcal{S}) = [1/2, 1/4, 1/4]$. In general, this choice should reflect the relative frequency of time signatures. We propose the following form for the prior of $\mathcal{S} = [s_i]$

Table 1: $w(s_i)$

s_i	2	3	5	7	11	13	17	o/w
$w(s_i)$	0	1	2	3	4	5	6	∞

$$p(\mathcal{S}) \propto e^{-\xi \sum_i w(s_i)} \tag{16}$$

where $w(s_i)$ is a simple weighting function given in Table 1. This form prefers subdivisions by small prime numbers, which reflects the intuition that rhythmic subdivisions by prime numbers such as 7 or 11 are far less common then subdivisions such as 2 or 3. The parameter ξ distributes probability mass over the primes. When $\xi = 0$, all subdivision schemata are equally probable. As $\xi \to \infty$, only subdivisions with $s_i = 2$ have non-zero probability.

4 Verification of the Model

To choose the likelihood $p(\mathbf{t}|\mathbf{c})$ and the prior $p(\mathbf{c})$ in a way which is perceptually meaningful, we analyzed data obtained from an psychoacoustical experiment where ten well trained subjects (nine conservatory students and a conservatory professor) have participated Desain et al. (1999). The experiment consisted of a perception task and a production task.

4.1 Perception Task

In the perception task the subjects were asked to transcribe 91 different *stimuli*. These rhythms consisted of four onsets $t_0 \dots t_3$ where t_0 and t_3 were fixed and occur exactly on the beat (Figure 10). First a beat is provided to subjects (count in), and then the stimulus is repeated 3 times with an empty bar between each repetition. Subjects were allowed to use any notation as a response and listen to the stimulus as much as they wanted. In total, subjects used 125 different notations, from which 57 were used only once and 42 are used more than three times. An example is depicted in Figure 11(a). From this data, we estimate the posterior as

$$q(\mathbf{c}_j|\mathbf{t}_k) = n_k(\mathbf{c}_j)/\sum_j n_k(\mathbf{c}_j)$$

where $n_k(\mathbf{c}_j)$ denotes the number of times the stimulus $\mathbf{t}_k$ is associated with the notation $\mathbf{c}_j$.

4.2 Production Task

In the production task the subjects are asked to perform the rhythms that they have notated in the perception task. An example is shown in Figure 11(a). For each notation $\mathbf{c}_j$ we assume a gaussian distribution where

$$\hat{q}(\mathbf{t}|\mathbf{c}_j) = \mathcal{N}(\mu_j, \Sigma_j) \tag{17}$$

The mean and the covariance matrix are estimated from production data by

$$\mu_j = \frac{1}{N_j}\sum_k \mathbf{t}_{k,j} \tag{18}$$

$$\Sigma_j = \frac{1}{N_j - 1}\sum_{k,l}(\mathbf{t}_{k,j} - \mu_j)(\mathbf{t}_{l,j} - \mu_j)^T \tag{19}$$

where $\mathbf{t}_{k,j}$ is the k'th performance of $\mathbf{c}_j$ and N_j is the total count of these performances in the data set. In Section 3.5.1 we proposed a model in which the correlation between two onset decreases with increasing inter-onset interval. The correlation coefficient and the estimated error bars are depicted in Figure 12, where we observe that the correlation decreases with increasing distance between onsets.

4.3 Estimation of model parameters

The probabilistic model $p(\mathbf{c}|\mathbf{t})$ described in the previous section can be fitted by minimizing the “distance”to the estimated target $q(\mathbf{c}|\mathbf{t})$. A well known distance measure between two probability distributions is the Kullback-Leiber divergence (Cover and Thomas, 1991) which is given as

$$\mathrm{KL}(q||p) = \int d\mathrm{x} q(\mathrm{x}) \log \frac{q(\mathrm{x})}{p(\mathrm{x})} \tag{20}$$

The integration is replaced by summation for discrete probability distributions. It can be shown (Cover and Thomas, 1991) that $\mathrm{KL}(q||p) \geq 0$ for any q, p, and vanishes if and only if $q = p$.

The KL divergence is an appropriate measure for the rhythm quantization problem. We observe that for many stimuli, subjects give different responses and consequently it is difficult to choose just one “correct” notation for a particular stimulus. In other words, the target distribution $q(\mathbf{c}|\mathbf{t})$ has its mass distributed among several codevectors. By minimizing the KL divergence one can approximate the posterior distribution by preserving this intrinsic uncertainty.

The optimization problem for the perception task can be set as

$$\begin{array}{ll} \min. & \mathrm{KL}(q(\mathbf{c}|\mathbf{t})s(\mathbf{t})||p(\mathbf{c}|\mathbf{t})s(\mathbf{t})) \\ \text{s.t.} & \sigma > 0 \\ & -1 < \eta < 1 \\ & \lambda, \xi, \gamma \ \ \text{unconstrained} \end{array} \tag{21}$$

where $s(\mathbf{t}) \propto \sum_k \delta(\mathbf{t} - \mathbf{t}_k)$ is the distribution of the stimuli. This is a distribution, which has positive mass only on the stimuli points $\mathbf{t}_k$. This measure forces the model to fit the estimated posterior at each stimulus point $\mathbf{t}_k$. We note that

$$p(\mathbf{c}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{c};\sigma,\lambda,\eta)p(\mathbf{c};\xi,\gamma)}{\sum_{\mathbf{c}} p(\mathbf{t}|\mathbf{c};\sigma,\lambda,\eta)p(\mathbf{c};\xi,\gamma)} \tag{22}$$

This is in general a rather difficult optimization problem due to the presence of the denominator. Nevertheless, since the model has only five free parameters, we were able to minimize Eq. 21 by a standard BFGS Quasi-Newton algorithm (MATLAB function fminu). In our simulations, we observed that the objective function is rather smooth and the optimum found is not sensitive to starting conditions, which suggests that there are not many local minima present.

4.4 Results

The model is trained on a subset of the perception data by minimizing Eq. 21. In the training, we used 112 different notations (out of 125 that the subjects used in total), which could be generated by one of the subdivision schemas in Table 2. To identify the relative importance of model parameters, we optimized Eq. 21 by clamping some parameters. We use a labeling of different models as follows: Model-I is the “complete” model, where all parameters are unclamped. Model-II is an onset quantizer ($\Sigma = \sigma^2\mathbf{I}$), where only prior parameters are active. Model-III is (almost) an IOI quantizer where the correlation between onsets is taken to be $\rho = 0.98$. Model-IV is similar to Model I with the simplification that the covariance matrix is constant for all codevectors. Since $\lambda = 0$, $\rho = \eta$. Model-V is an onset quantizer with a flat prior, similar to the quantizers used in commercial notation packages and Model-VI has only the performance model parameters active.

In Model-VII, the parameters of the performance model $p(\mathbf{t}|\mathbf{c})$ are estimated from the production data. The model is fitted to the production data $\hat{q}$ by minimizing

$$\mathrm{KL}(\hat{q}(\mathbf{t}|\mathbf{c})q(\mathbf{c})||p(\mathbf{t}|\mathbf{c})q(\mathbf{c})) \tag{23}$$

where $q(\mathbf{c}_j) = \sum_k n_k(\mathbf{c}_j) / \sum_{k,j} n_k(\mathbf{c}_j)$, i.e. a histogram obtained by counting the subject responses in the perception experiment.

Although approximating the posterior at stimuli points is our objective in the optimization, for automatic transcription we are also interested into the classification performance. At each stimuli $\mathbf{t}_k$, if we select the response which the subjects have chosen the most, i.e. $\mathbf{c}_k^* = \arg\max_{\mathbf{c}} q(\mathbf{c}|\mathbf{t}_k)$, we can achieve maximum possible classification rate on this dataset, which is given as

$$\text{CR}_{\text{Target}} = \frac{n_k(\mathbf{c}_k^*)}{Z} \times 100 \tag{24}$$

Here, $Z = \sum_{k,\mathbf{c}} n_k(\mathbf{c}_k^*)$, the total number of measurements. Similarly, if we select the codevector with the highest predicted posterior $\mathbf{c}_k^* = \arg\max_{\mathbf{c}} p(\mathbf{c}|\mathbf{t}_k)$ at each stimulus, we achieve the classification rate of the Model denoted as CR_{Model}. The results are shown in Table 3. The clamped parameters are tagged with an '=' sign. The results are for a codebook consisting of 112 codevectors, which the subjects have used in their responses and could have been generated by one of the subdivisions in Table 2.

i	$\mathcal{S}_i$
1	[2, 2, 2, 2]
2	[3, 2, 2]
3	[3, 3, 2]
4	[5, 2]
5	[7, 2]
6	[11]
7	[13]
8	[5, 3]
9	[17]
10	[7, 3]

Table 2: Subdivisions

Model	Prior		Likelihood			Results	
Label	ξ	γ	σ	λ	η	KL	$\text{CR}_{\text{Model}}/\text{CR}_{\text{Target}}$
I	1.35	0.75	0.083	2.57	0.66	1.30	77.1
II	1.34	0.75	0.086	= 0	= 0	1.41	71.3
III	1.33	0.77	0.409	= 0	= 0.98	1.96	51.4
IV	1.34	0.74	0.084	= 0	0.39	1.34	75.3
V	= 0	= 0	0.085	= 0	= 0	1.92	29.7
VI	= 0	= 0	0.083	2.54	0.66	1.89	32.7
VII	1.43	0.79	! 0.053	! 3.07	! 0.83	1.89	84.3

Table 3: Optimization Results. $\text{CR}_{\text{Target}} = 48.0$. Values tagged with a '=' are fixed during optimization. Values estimated from the production experiment are tagged with a '!'. The meanings of the columns are explained in the text.

Model-I performs the best in terms of the KL divergence, however the marginal benefit obtained by choosing a correlation structure, which decreases with increasing onset distances (obtained by varying λ) is rather small. One can achieve almost the same performance by having a constant correlation between onsets (Model-IV). By comparing Model-IV to Models II and

III, we can say that under the given prior distribution the subjects are employing a quantization strategy, which is somehow between a pure onset quantization and IOI-quantization. The choice of the prior is very important which can be seen from the results of Model-V and Model-VI, which perform poor due to the flat prior assumption.

Model-VII suggests that for this data set (under the assumption that our model is correct) the perception and production processes are different. This is mainly due to the spread parameter σ, which is smaller for the production data. The interpretation of this behavior is that subjects deviate less from the mechanical mean in a performance situation. However, this might be due to the fact that performances were carried out in lack of any context, which forces the subjects to concentrate on exact timing. It is interesting to note that almost the same correlation structure is reserved in both experiments. This suggests that there is some relation between the production and perception process. The classification performance of Model-VII is surprisingly high; it predicts the winner accurately. However the prediction of the posterior is poor, which can be seen by the high KL divergence score.

For visualization of the results we employ an interpolation procedure to estimate the target posterior at other points than the stimuli (See Appendix A). The rhythm space can be tiled into regions of rhythms, which are quantized to the same codevector. Estimated tiles from experimental data are depicted in Figure 13(a).

In practice, it is not feasible to identify explicitly a subset of all possible codevectors, which have non-zero prior probability. For example, the number of notations which can be generated by subdivisions in Table 2 is 886 whereas the subjects used only 112 of these as a response. This subset must be predicted by the model as well. A simple grid quantizer tries to approximate this subset by assigning a constant prior probability to codevectors only up to a certain threshold depth. The proposed prior model can be contrasted to this schema in that it distributes the probability mass in a perceptually more realistic manner. To visualize this, we generated a codebook consisting of all 886 codevectors. The tilings generated by Model-I and Model-V for this codebook are depicted in Figure 13(b) and 13(c). To compare the tilings, we estimate the ratio

$$\text{Match} = \frac{A_{\text{match}}}{A_{\text{total}}} \times 100 \qquad (25)$$

where A_{match} is the area where the model matches with the target and A_{total} is the total area of the triangle. Note that this is just a crude approximation to the classification performance under the assumption that all rhythms are equally probable. The results are shown in Table. 4.

	I	II	III	IV	V	VI	VII
Match	58.8	53.5	36.1	59.0	3.8	3.1	56.7

Table 4: Amount of match between tilings generated by the target and models

5 Discussion and Conclusion

In this article, we developed a vector quantizer for transcription of musical performances. We considered the problem in the framework of Bayesian statistics where we proposed a quantizer model. Experimentally, we observe that even for quantization of simple rhythms, well trained subjects give quite different answers, i.e. in many cases, there is not only one correct notation. In this respect, probabilistic modeling provides a natural framework. The model is verified and optimized by data obtained from the psychoacoustical experiment. The optimization results suggest that prior and likelihood parameters are almost independent, since clamping one set of parameters affects the optimal values of others only very slightly. This property makes the interpretation of the model easier.

It is important to note that in the derivations we did not used any other attributes of notes (e.g. duration, pitch), which give additional information for better quantization. Another point is that in quantization of real performances, context information plays also an important role. The main advantage of Bayesian framework is that all such improvements can be integrated by modifying the likelihood and prior distributions suitably. As already demonstrated, since all the assumptions are stated as distributions, corresponding optimal parameters can be estimated from experimental data.

Acknowledgements

This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Dutch Ministry of Economic Affairs. The first author is thankful to David Barber for stimulating discussions.

A Estimation of the posterior from subject responses

Let $\mathbf{t}_k$ be the stimuli points. The histogram estimate at $\mathbf{t}_k$ is denoted by $q(\mathbf{c}_j|\mathbf{t}_k)$. We define a kernel

$$G(\mathbf{t}; \mathbf{t}_0, \sigma) = \exp(-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{t}_0\|^2) \tag{26}$$

where $\|\mathrm{x}\|$ is the length of the vector x. Then the posterior probability of $\mathbf{c}_j$ at an arbitrary point $\mathbf{t}$ is given as

$$q(\mathbf{c}_j|\mathbf{t}) = \sum_k \alpha_k(\mathbf{t}) q(\mathbf{c}_j|\mathbf{t}_i) \tag{27}$$

where $\alpha_k(\mathbf{t}) = \frac{G(\mathbf{t};\mathbf{t}_k,\sigma)}{\sum_r G(\mathbf{t};\mathbf{t}_r,\sigma)}$. We have taken $\sigma = 0.04$.

References

Carlos Agon, Gérard Assayag, Joshua Fineberg, and Camilo Rueda. Kant: A critique of pure quantification. In *Proceedings of the International Computer Music Conference*, pages 52–9, Aarhus, Denmark, 1994. International Computer Music Association.

E. F. Clarke. Structure and expression in rhythmic performance. In P. Howell, I. Cross, and R. West, editors, *Musical structure and cognition*. Academic Press, Inc., London, 1985.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

P. Desain, R. Aarts, A. T. Cemgil, B. Kappen, H. van Thienen, and P. Trilsbeek. Robust time-quantization for music. In *Proceedings of the AES 106th Convention*, page (in submission), Munich, Germany, May 1999. AES.

P. Desain and H. Honing. Quantization of musical time: a connectionist approach. In P. M. Todd and D. G. Loy, editors, *Music and Connectionism.*, pages 150–167. MIT Press., Cambridge, Mass, 1991.

P. Desain, H. Honing, and K. de Rijk. The quantization of musical time: a connectionist approach. In *Music, Mind and Machine: Studies in Computer Music, Music Cognition and Artificial Intelligence*, pages 59–78. Thesis Publishers, Amsterdam, 1992.

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, Inc., New York, 1973.

E. W. Large. Beat tracking with a nonlinear oscillator. In *Working Notes of the IJCAI Workshop on AI and Music*, pages 24–31. International Joint Conferences on Artifical Intelligence, 1995.

H. Christopher Longuet-Higgins. *Mental Processes: Studies in Cognitive Science*. 1987. 424p.

J. Pressing and P. Lawrence. Transcribe: A comprehensive autotranscription program. In *Proceedings of the International Computer Music Conference*, pages 343–345, Tokyo, 1993. Computer Music Association.

P. Toiviainen. An interactive midi accompanist. *Computer Music Journal*, 22:4:63–75, 1999.

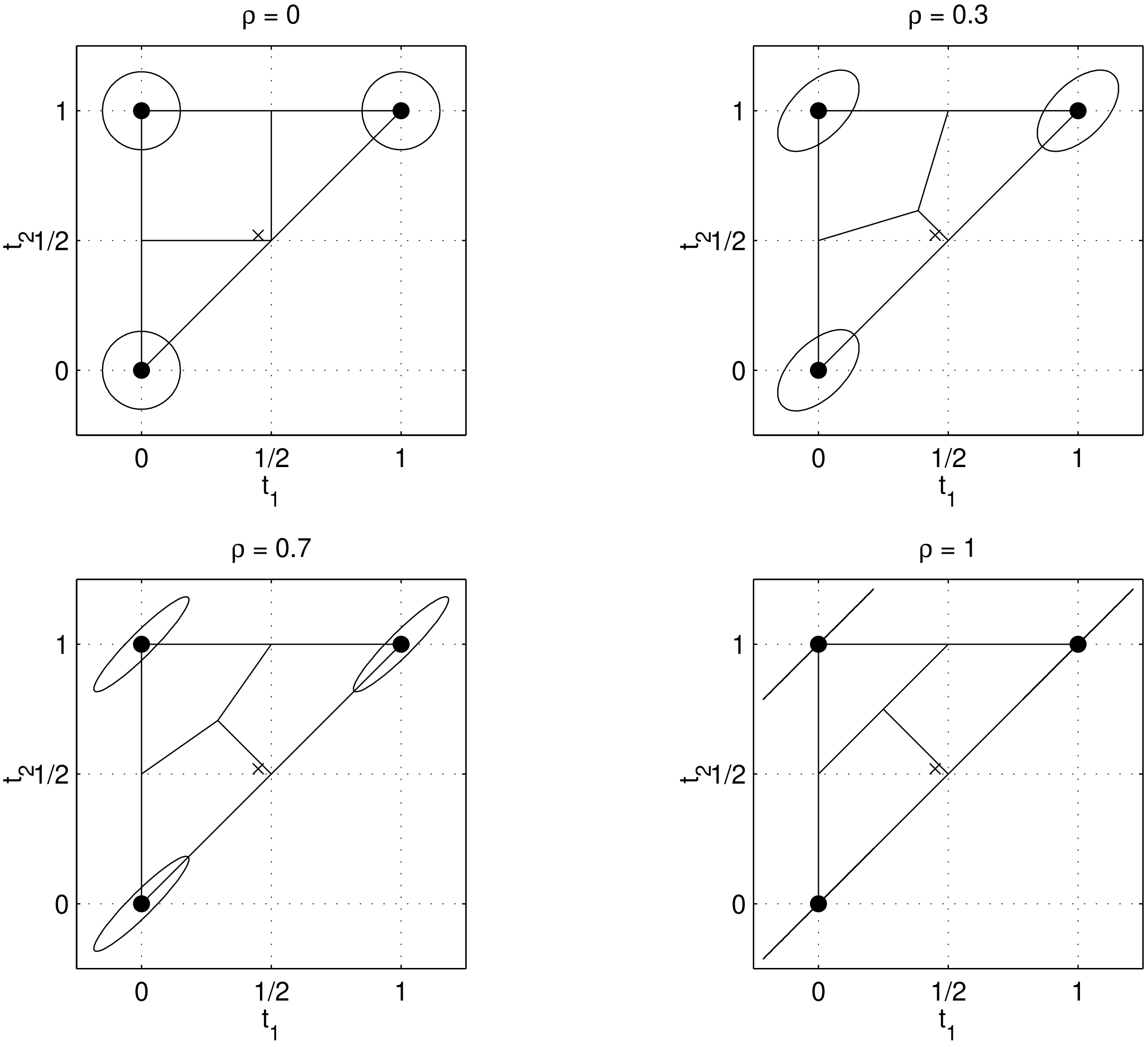
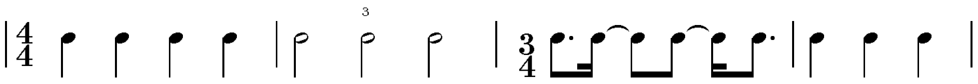


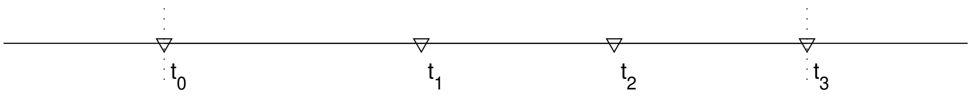
Figure 8: Tiling for choices of ρ and constant $p(\mathbf{c})$. Onset quantization (i.e. grid quantization) used by many commercial notation packages corresponds to the case where $\rho = 0$. IOI quantization appears when $\rho \to 1$. Note that different correlation structures imply different quantization decisions, not necessarily onset- or IOI-quantization. The cross corresponds to the rhythm $\mathbf{t} = [0.45, 0.52]$.

(a) In lack of any other context, both onset sequences will sound the same. However the first notation is more complex

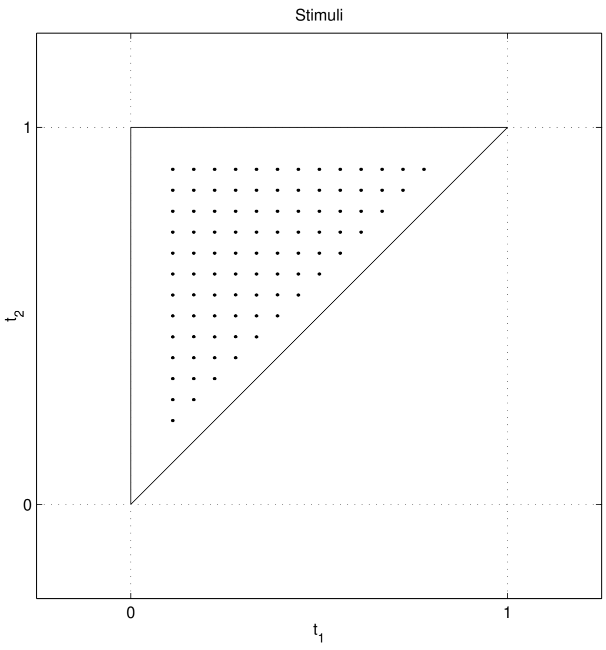


(b) Assumed time signature determines the complexity of a notation

Figure 9: Complexity of a notation

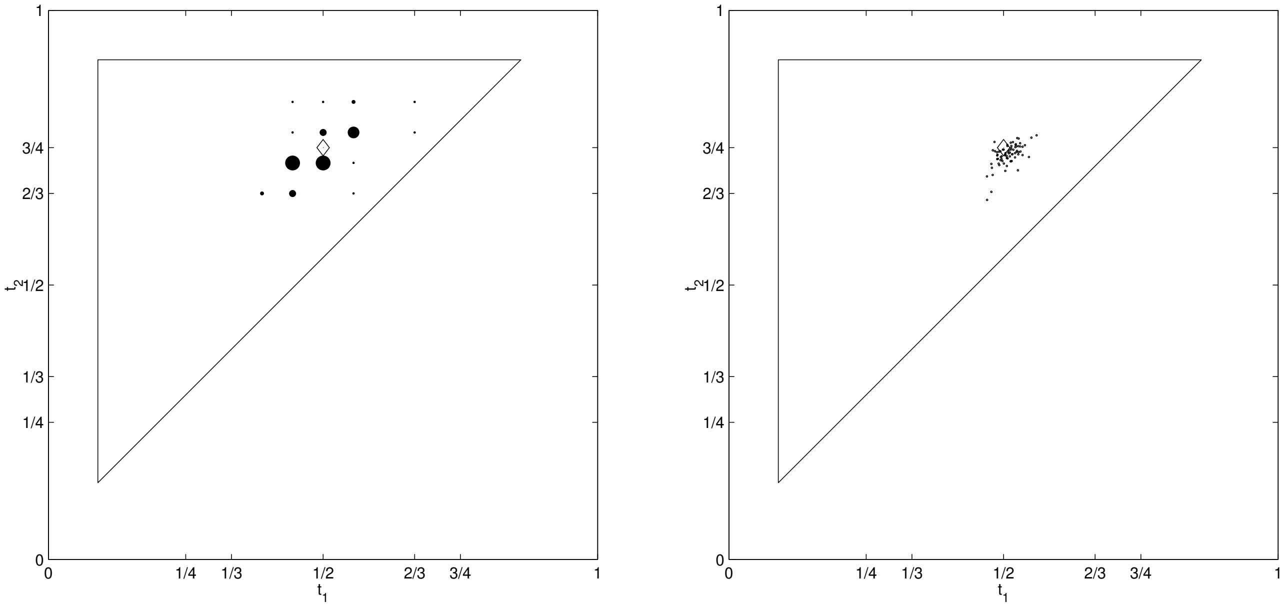


(a) Stimulus



(b) Stimuli for the perception experiment. The dots denote the rhythms $\mathbf{t}_k$, where $k = 1 \ldots 91$. Grid spacing is 56ms.

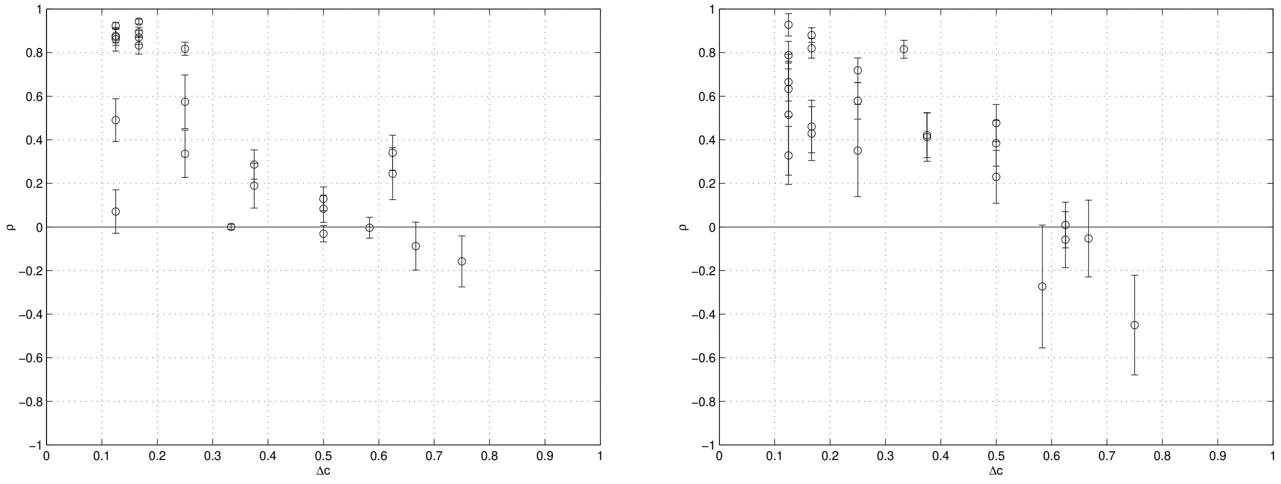
Figure 10: Stimulus of the Perception Task



(a) Perception

(b) Production

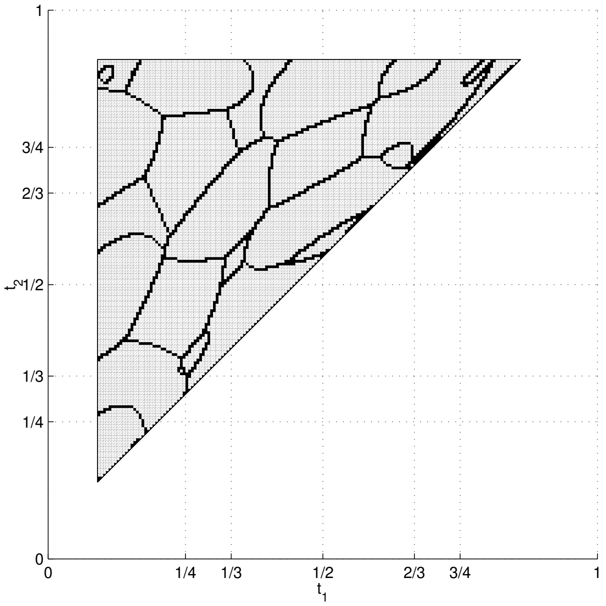
Figure 11: Perception and Production of the rhythm [2 1 1] (**c** = [0.5 0.75]). The diamond corresponds to the mechanical performance. In 11(a), the size of the circles is proportional to the estimated posterior $q(\mathbf{c}_j|\mathbf{t}_k)$. In 11(b), the dots correspond to performances of the rhythm.



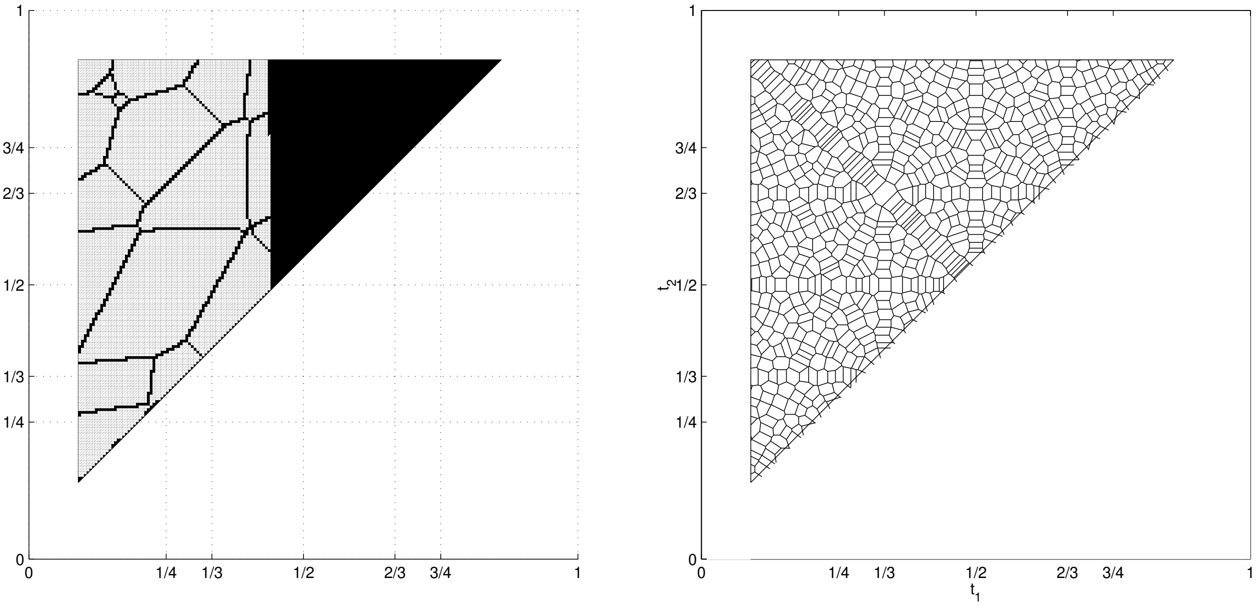
(a) Production

(b) Perception

Figure 12: Estimated correlation coefficient as a function of $\Delta c = (c_2 - c_1)$ on all subject responses.



(a) Target



(b) Model-I: $(\xi, \gamma, \sigma, \lambda, \eta) = (1.35, 0.75, 0.083, 2.57, 0.66)$

(c) Model-V: $(\xi, \gamma, \sigma, \lambda, \eta) = (0, 0, 0.085, 0, 0)$

Figure 13: Tilings of the rhythm space by $\mathbf{c}^* = \operatorname*{argmax}_{\mathbf{c}} p(\mathbf{c}|\mathbf{t})$. The tiles denote the sets of rhythms, which would be quantized to the same codevector. Both Model-I and Model-V use the same codebook of 886 codevectors. Since Model-V assigns the same prior probability to all codevectors, the best codevector is always the nearest codevector (in Euclidian distance) and consequently the rhythm space is highly fragmented.