A Bayesian Technique for Distinguishing Between Melody Notes and Grace Notes in Recorded Performances

W. James MacLean
Edward S. Rogers Sr. Department of Electrical & Computer Engineering,
University of Toronto,
Toronto, Canada, M5S 3G4
maclean@eecg.toronto.edu

Abstract

This paper describes a Bayesian technique for distinguishing melody notes from grace notes in a recorded performance. A mixture model is used to describe probabilities of note durations, and a graphical model relates contextual information (the pitches of three successive notes). Practical examples based on MIDI-encoded bagpipe performances are given.

1 Introduction

A *gracenote* can be defined as a note of short duration whose purpose is to provide punctuation. The duration of a gracenote is not included when summing the time in a bar and in performance the duration will be unrelated to the progression defined by eighth notes, quarter notes and so on. Gracenotes are typeset differently from melody notes, typically being set in a much smaller type size. For many instruments (*e.g.* piano) they are used at the discretion of the composer and are not a necessity, whereas for some instruments (*e.g.* bagpipe) they are absolutely essential to the production of melodies.

Being able to automatically detect gracenotes in performances has a number of possible uses. Perhaps the most obvious is in music transcription systems [11]. Here software is used to transform a performance into a typeset score without requiring the musician to manually enter the notes. In a piece that contained gracenotes the system would need to be able to distinguish these so they could be typeset appropriately. Another possible use involves attempts to analyze the rhythmic structure in performance using computers [9, 8, 10, 7]. Although the duration of gracenotes is ignored in the scoring process, they cannot be ignored in performance and form an integral part of the rhythmic framework of the melody. Distinguishing between gracenotes and melody notes in performance is therefore a matter of concern. Yet another use can be imagined in the realm of music education. If a computer could analyze performance in real-time, it could be used to monitor the practising of a music student and provide immediate feedback on errors (or even reward good technique and style) thus minimizing the amount of time spent practising wrongly.

A simple technique would be to observe the durations of notes in a performance, and segment the gracenotes by identifying the note population with the shortest duration as being gracenotes. This simple scheme works reasonably well when the fastest melody notes are long in comparison to the gracenotes, but fails in pieces played at fast tempos. In this case it is necessary to include additional (contextual) information, such as the pitch of the note and even adjacent notes, in order to correctly label gracenotes. This paper presents such a method. In particular, a Bayesian formulation is given in which note durations are clustered using the EM algorithm and related to pitch via a graphical model. Results are given for bagpipe performances, although the method is not instrument-specific.

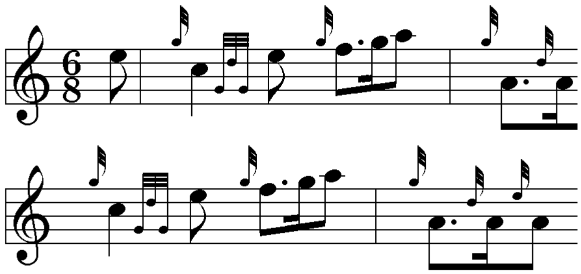

Figure 1: An extract from a typical bagpipe score is shown. Note the presence of single and multiple gracenote groupings. All melody notes are scored with stems down to facilitate the inclusion of gracenotes.

2 Methods

In this paper results will be presented using music performed on a *Great Highland Bagpipe* (GHB). The GHB has a nine-note scale spanning just over an octave. It relies on gracenotes for melody production as it is impossible to sound the same note twice in succession without one. Also, since the GHB cannot produce a rest or change its volume, the only means left for expression is that provided by rhythm. Gracenotes are no longer just punctuation required to separate notes of the same pitch. Instead, single gracenotes and multi-gracenote groupings are used to enhance rhythm and give texture to the music. A couple of bars extracted from a typical score are shown in Figure 1. Further details can be found in [2]. This paper will limit its scope to the detection of single gracenotes[1].

2.1 Clustering Based on Note Duration

A histogram of note durations from a musical performance will contain mutliple peaks, one for each note duration found in the score. Not all notes of a class (*e.g.* eighth notes) will have exactly the same duration. Rather, we expect a distribution of durations centred around some mean value. The histogram of all note durations would be expected then to contain multiple peaks or *clusters*, each one corresponding to a particular duration class. The locations of the clusters will be related. One might expect the mean of the quarter-note peak, for example, to be twice that for eighth notes. Cut quarter-notes would coincide with the eighth-note's peak, and dotted quarter-notes would have a peak of their own. Gracenotes would be expected to have the shortest duration, and the mean of this cluster is not expected to be related to that of the timed notes. The individual clusters can each be modelled as Gaussians.

A *mixture model* [6] is a useful statistical model for distributions with multiple peaks. It can be used to describe the histogram of note durations with the sum of individual distributions:

$$p(d) = \sum_{i=0}^{N} \pi_i p_i(d|\mu_i, \sigma_i)$$

where d is the duration of a note, and $p(d)$ is the probability density for observing a note of that duration. The function $p_i(d)$ is the probability density for observing a note of duration d given that it comes from the i-th peak. For example, $p_0(d)$ might represent the probability density that a gracenote of duration d is observed, $p_1(d)$ that a sixteenth-note of duration d is observed, and so on. Each distribution has its own mean (μ_i) and variance (σ_i^2). The total number of peaks is $N+1$, and the π_i are chosen so that they are each positive and sum to unity, thus providing normalizing factors for the overall distribution. We can think of π_0 as being the *a priori* probability of any given note being a gracenote before knowing its duration. Likewise π_1 gives the probability of a sixteenth-note and so on.

We can enforce a relationship between the μ_i's to represent our expectation that eighth-notes will approximately twice as long as sixteenth notes by setting $\mu_2 = 2\mu_1$. Temporarily ignoring dotted notes, we can continue by setting $\mu_i = 2\mu_{i-1}$ for all distributions except the gracenotes. If we likewise link the σ's, we reduce the number of parameters in the model to four: μ_0, σ_0, μ_1 and σ_1. This leads to

$$p(d) = \pi_0 p(d|\mu_0, \sigma_0) + \sum_{i=1}^{N} \pi_i p_i(d|2^{i-1}\mu_1, 2^{i-1}\sigma_i) \ .$$

Since we do not know in advance which notes belong to which clusters, we need a technique to simul-

[1] The detection of multiple successive gracenotes turns out to be a simpler problem as these groups are stereotypical and quite distinctive

taneously segment the notes and estimate the distribution parameters. The *EM Algorithm* [3] is one way to achieve this. It is a *Maximum Likelihood Estimation* (MLE) technique designed to operate with missing data. In our case, the complete data are the parameters for the mixture model togther with the knowledge of which observed durations belong to which $p_i(d)$. The latter is unknown, and constitutes the missing data. The method proceeds as follows:

1. Start with an initial estimate for the parameters. If we know the tempo of the performance and the shortest note class we can intelligently guess what μ_1 might be, for example. If we assume that all note durations are equi-probable, we set each of the π_i to $1/N + 1$.

2. Using the parameters, estimate the likelihood that the j-th observed duration comes from cluster i via

$$l_{ij} = \frac{\pi_i p(d_j|\mu_i, \sigma_i)}{\sum_{k=1}^{N} \pi_k p(d_j|\mu_k, \sigma_k)} \quad .$$

   This is called the "E-step" (expectation).

3. Using the l_{ij} from the previous step, recompute the distribution parameters. For example, the means can be computed as

$$\mu_i' = \frac{\sum_{j=1}^{m} l_{ij} d_j}{\sum_{j=1}^{m} l_{ij}}$$

   where m is the number of notes observed. The mixing parameters can be re-estimated as

$$\pi_i' = \frac{\sum_{j=1}^{m} l_{ij}}{m} \quad .$$

   This is the "M-step" (maximization).

Steps 2 and 3 are repeated until the parameters converge. Upon completion of the algorithm the l_{ij} give a soft-segmentation of durations with respect to the different clusters, and the distribution parameters are known. It must be pointed out that the problem is non-linear, and poor initial estimates for the parameters can lead to specious results. Despite this problem, the method has proven useful in fitting mixture distributions to observed note durations.

2.2 Use of Contextual Information

For many pieces we expect only minimal overlap between the tails of the component distributions. However, if we consider pieces with fast tempos (*e.g.* jigs, reels and other dance music) we may find that the distribution representing the shortest melody notes may overlap significantly with the distribution for gracenotes. As already mentioned, the latter tend to be played "ballistically", often as fast as the performer is capable of and certainly without regard to the durations of other notes. In this event note duration alone is insufficient to label notes as gracenotes or melody notes. We require more information.

The other piece of information typically available is the pitch of each note. This is a rich source of contextual information as not all note combinations are equally likely. For example, in bagpipe music, two successive melody notes having the same pitch *must* be separated by a gracenote. Therefore if we encounter three notes where the first and last have the same pitch and the middle note is of short duration, the middle note has a higher probability of being a gracenote. For any given instrument not all note-gracenote combinations may be possible given technical considerations. Among those combinations that are possible, some may be uncommon due to aesthetic reasons or due to being technically too demanding.

The approach we will take is to examine three note groupings with the intent of answering the question "Is the middle note a gracenote?" We will limit ourselves to the case of single gracenotes punctuating melody notes. It is assumed that the pitch and duration of each note is known. Let n_1, n_2 and n_3 represent successive note pitches, and d_1, d_2 and d_3 their durations respectively. Further, let G be a Boolean-valued variable that is true if the second note is a gracenote. If we knew the joint probability density for these variables, $p(n_1, n_2, n_3, d_1, d_2, d_3, G)$ we could use it to provide context for determining whether or not the second note is a gracenote given observations for the first six values.

A probability density with seven variables might be expected to be difficult to estimate. It is possible to make the problem more tractable by imposing some structure on this function. For example, we might expect the durations to be independent of the

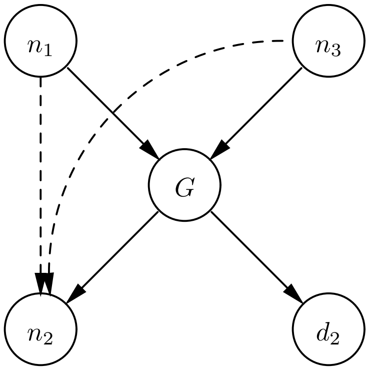

Figure 2: This figure shows a graphical model relating the variables in the probability density for gracenote labelling.

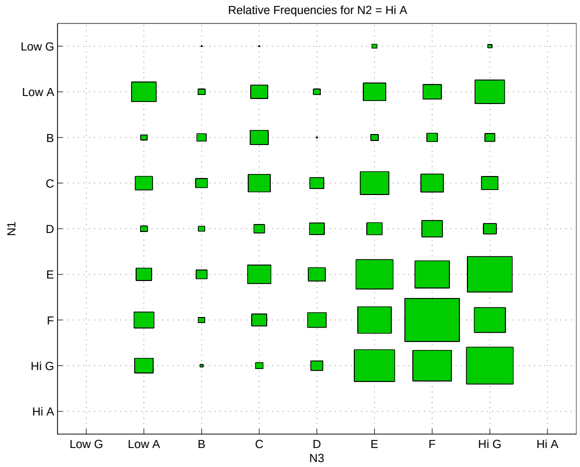

Figure 3: The distribution $p(n_2|n_1, n_3, G)$ is given for $n_2 =$ Hi A and $G =$ true. Larger boxes indicate larger probability, and absence of a box indicates 0 probability (for example, for the $n_1 =$ Hi A row and $n_3 =$ Hi A column we have zeros since we require $n_1 \neq n_2$ and $n_2 \neq n_3$).

pitches and only depend on whether or not the note is a gracenote. In this case d_1 and d_3 become immaterial, and d_2 only depends on G. Further, we might expect that p_1 and p_3 are independent of one another. We cannot assume that p_2 is independent of either p_1 or p_3 since, at the very least, no two successive notes can have the same pitch. It is possible to use a graph to describe the structure of a density function [5, 4]. Figure 2 shows the model proposed for labelling gracenotes.

In the model, the values for n_1 and n_3 are assumed to be independent of all other variables. The value of n_2 is related to both n_1 and n_3 as well as G. The value of G is also affected by n_1 and n_3 as not all combinations of n_1 and n_3 will permit a gracenote in between, but if $n_1 = n_3$ then the likelihood of G being true increases. Finally, once the value of G is known we assume the value of d_2 is not affected by any of the pitch values. The density $p(d_2|G)$ is easily computed from the mixture model outlined in Section 2.1. Based on the proposed model, we can factor the density as

$$\begin{aligned} & p(n_1, n_2, n_3, d_2, G) \\ & = p(n_1)p(n_3)p(G|n_1, n_3)p(n_2|G, n_1, n_3)p(d_2|G) \quad . \end{aligned} \tag{1}$$

The improtant result here is that we can now estimate the individual factors separately. The densities $p(n_1) = p(n_3)$ can be estimated by observing the frequencies of melody notes of different pitch. All but the last factor can be estimated from scores since they do not involve d_2, and can therefore be done offline in advance.

We are interested in the quantity $p(G|n_1, n_2, n_3, d_2)$. Using Bayes law we write

$$\begin{aligned} & p(G|n_1, n_2, n_3, d_2) \\ & = \frac{p(n_1,n_2,n_3,d_2,G)}{p(n_1,n_2,n_3,d_2,G)+p(n_1,n_2,n_3,d_2,\bar{G})} \\ & = \frac{p(G|n_1,n_3)p(d_2|G)p(n_2|n_1,n_3,G)}{p(G|n_1,n_3)p(d_2|G)p(n_2|n_1,n_3,G)+p(\bar{G}|n_1,n_3)p(d_2|\bar{G})p(n_2|n_1,n_3,\bar{G})} \end{aligned}$$

where the last line involves the substitution of Eq 1. The notation $\bar{G}$ is used as shorthand for $G =$ false and G for $G =$ true. The terms involving $p(d_2|G)$ and $p(d_2|\bar{G})$ are found from the mixture model for note durations. The terms involving $p(n_2|n_1, n_3, G)$ and $p(n_2|n_1, n_3, \bar{G})$ are found from non-parametric estimation using musical scores. An example of this distribution for Hi A gracenotes is given in Figure 3.

3 Results

In this section results are given from applying the techniques described in Section 2 to performance data. The data (note pitches and durations) were

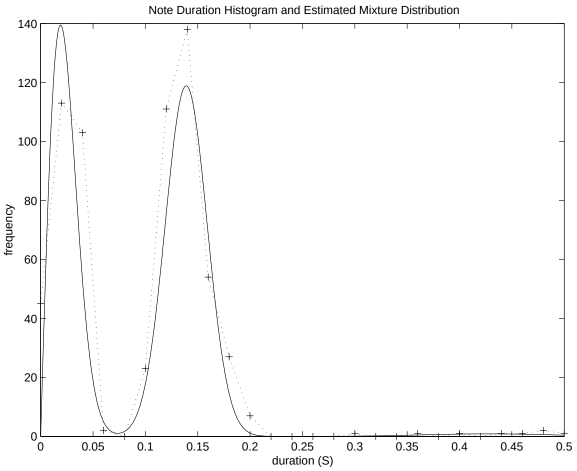

Figure 4: This figure shows the histogram of note durations from the jig "Duncan the Gauger". Also shown are the components of the fitted mixture distribution. The mixture density has been scaled by $m/\sum_i d_j$. While the height of the peaks do not match exactly, remember that the area underneath the peaks should match.

obtained using a DegerPipes[2] MIDI bagpipe. Performances were recorded and note durations and pitches were extracted into text files for subsequent import into Matlab. A small number of outliers existed in the MIDI file in the form of spurious notes of extremely short duration caused by innacurate fingerings by the performer (the author) or by capacitive delays in the sensors in the MIDI chanter.

The resulting histogram of note durations from the jig "Duncan the Gauger" is shown in Figure 4. A Rayleigh density was used to model the gracenote cluster in order to avoid the left tail of this density from going below zero.

The estimated mixture density is a good match for the histogram data, and all non-spurious notes are correctly labelled as either melody notes or gracenotes based on duration alone. However, the spurious notes are also labelled as gracenotes even though they cannot be based on their pitch values. In this performance all gracenotes are correctly labelled by the model.

Mixture Component	**Mixture Proportion**	**Density Parameters**
0	$\pi_0 = 0.418$	$\alpha_0 = 0.019$
1	$\pi_1 = 0.5693$	$\mu_1 = 0.139$ $\sigma_1 = 0.020$
2	$\pi_2 = 0.0127$	$\mu_2 = 0.427$ $\sigma_2 = 0.062$

Table 1: This table shows the estimated parameters for the model. The total number of notes is $m = 631$. Component 0 represents gracenotes, component 1 represents 1/8-th notes, and component 2 represents dotted 1/4-th notes. The mean duration for component 1 suggests the tempo was 144 bpm for this performance.

The estimation of the density $p(n_2|n_1, n_3, G)$ was achieved by analyzing scores directly. Fortunately there exists a large number of pipe tunes electronically typeset using software called "Bagpipe Music Writer"[3]. The underlying file format is text and was easily analyzed using a custom Perl script. A database of over 1700 tunes was used to create a non-parametric estimate of this density function. The dataset comprised 102,817 melody notes and 37,290 single gracenotes. It should be noted that some errors existed in the scores in the database, most often of the form of placing a gracenote adjacent to a melody note with the same pitch. The Perl scripts included code to report such errors and ignore their associated data.

Table 2 shows the labellings for individual notes in the recorded performance. While in most cases $p(G|d)$ and $p(G|n_1, n_2, n_3, d)$ agree, the latter correctly detects artifacts caused by mis-fingering. In line 5 of the table we see a C between a B and a D. It has an extremely short duration (4 mS). Although it would seem to be a gracenote it is not. In changing from B to D on the GHB it is necessary to lift two fingers at the same time. In this case the performer has lifted one finger slightly ahead of the other, causing a 'false' C to be sounded (this would not normally be detectable on the acoustic version of the instrument, but the metal sensing pads of the DegerPipe

[2] Designed and built by Manfred Deger, see http://www.deger.de/.

[3] See http://home.istar.ca/˜rmm/ .

d (S)	p_0	p_1	p_2	P	Note
0.016	1.000	0.000	0.000	1.000	Lo G
0.133	0.000	1.000	0.000	0.000	B
0.038	1.000	0.000	0.000	1.000	Hi G
0.154	0.000	1.000	0.000	0.000	B
0.004	1.000	0.000	0.000	0.000	C
0.159	0.000	1.000	0.000	0.000	D
0.308	0.000	0.000	1.000	0.000	E
0.033	1.000	0.000	0.000	1.000	Hi G
1.300	0.000	0.000	0.000	0.000	Lo A
0.004	1.000	0.000	0.000	0.000	F
1.054	0.000	0.000	0.000	0.000	Hi A

Table 2: This table shows the output labelling of melody/gracenotes in “Duncan the Gauger”. The second column indicates the probability that the note is a gracenote based on duration alone. The fifth column is the *a posteriori* probability including contextual information. In most cases they agree, with the exception of the 5th line, where the note is in fact an artifact induced by mis-fingering: while it has a short duration, it is not in fact a gracenote. The last column gives the pitch-name of each note.

are able to detect slight differences in the timing of the movements of the two fingers and issue two MIDI note messages instead of one).

Results from a second example, this time a reel named “Colonel MacLeod”, are shown in Figure 5 and Tables 3 and 4. In addition to finding mis-fingerings, the model also correctly labelled short cut-eighth notes even though the duration-only model gave them high probability of being gracenotes. There are also instances of long gracenotes being given a higher posterior probability, thus helping better distinguish them from short cut-eighth notes. In this tune, excluding mis-fingerings, the duration-only method mislabelled 3 melody notes as gracenotes. The posterior correctly labelled all gracenotes and melody notes. There were a total of 330 non-spurious notes, 97 of which were gracenotes.

A weakness in the model appears in the form of back-to-back gracenotes (which violate the premise of a single-gracenote tune). In determining whether n_2 is a gracenote, the model completely ignores the labelling previously assigned to n_1. In most cases this has not proven problematic, but it does occur in examples not shown here. An appropriate solution to this problem would be the use of a Markov chain to consider the labellings of n_1 and n_3 when determining the correct label for n_2. This will increase the computational cost of the labelling as this far we have only used local operations.

In performing the EM clustering of note durations it was sometimes necessary to place upper (and lower) limits on the values of σ_i and α_0. The problem was particularly pronounced with the Rayleigh (gracenote) distribution: given the opportunity it would attempt to inflate its variance in order to acquire the data owned by the nearest population of melody notes. A similar behaviour can occur in which a process sets its mean to match a single data point and shrinks its variance to zero, in essence becoming a “grandmother cell”. This behaviour is well known [1] and selection of appropriate limits for variance values becomes an important issue. The idea of yoking all variances and means for melody notes (suggested in Section 2.1) is an attempt to deal with this issue. Unfortunately the variance of the gracenote population must be dealt with separately.

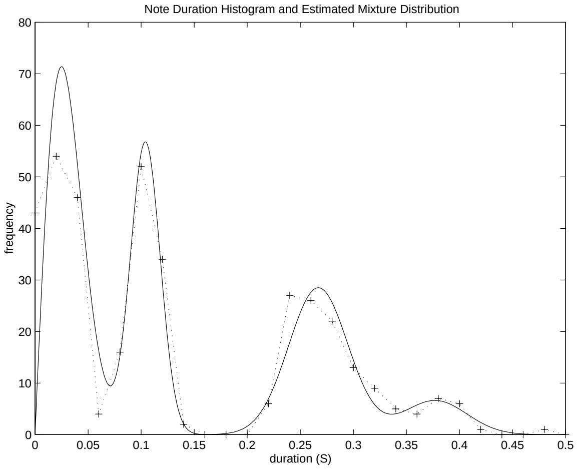

Figure 5: This figure shows the histogram of note durations from the reel "Colonel MacLeod". There are 4 components to the mixture in this case: gracenotes, cut-eighth notes, eighth notes and dotted-eighth notes. In this case the variance of the gracenote distribution had to be given an upper limit to prevent it from taking over the cut-eighth note population.

Mixture Component	**Mixture Proportion**	**Density Parameters**
0	$\pi_0 = 0.396$	$\alpha_0 = 0.025$
1	$\pi_1 = 0.268$	$\mu_1 = 0.104$ $\sigma_1 = 0.014$
2	$\pi_2 = 0.269$	$\mu_2 = 0.267$ $\sigma_2 = 0.028$
3	$\pi_3 = 0.067$	$\mu_3 = 0.377$ $\sigma_3 = 0.030$

Table 3: The total number of notes is $m = 378$. Component 0 represents gracenotes, component 1 represents cut-eighth notes, component 2 represents dotted-eighth notes, and component 3 represents quarter notes. The mean duration for component 4 suggests the tempo was about 120 bpm for this performance.

d (S)	p_0	p_1	p_2	p_3	P	Note
0.029	1	0	0	0	1	Hi G
0.065	0.882	0.118	0	0	0	C
0.313	0	0	0.914	0.086	0	E
0.031	1	0	0	0	1	Hi G
0.365	0	0	0.009	0.991	0	Lo A
0.055	0.993	0.007	0	0	1	D
0.094	0.009	0.991	0	0	0	C

Table 4: Line two shows a short C (cut-eighth), originally classified as a gracenote in the first column, correctly labelled as a melody note by the posterior listed in the sixth column. The second last line shows a D-gracenote, already correctly identified using its duration, with a higher posterior that helps confirm the correct labelling.

4 Discussion

Although the graphical model explored in this paper is specific to a particular instrument, the method as a whole is not. Similar reasoning could be used to derive the appropriate density function structure for other instruments. Ideally, it should be possible to learn the structure of the model directly from observed data, thus making the method completely general. As pointed out in Section 3 the model should be extended to include a Markov random field in order to model the effect of the labelling of n_1 and n_3 on that of n_2.

It should be noted that the EM clustering of note durations can be applied to mutliple-gracenote clusters, and obviously can also be used to cluster note durations where gracenotes are not involved. In the future the method will be applied to performances recorded from acoustic (non-MIDI) instruments, where notes can be extracted using wavelet decomposition techniques as outlined in [13, 12]. Any attempt at rhythmic analysis in the presence of single- and multi-gracenote groupings will require the ability to identify these clusters.

It is also possible to further sub-label based on a gracenote's function. For example, on a pipe gracenotes are usually performed in one of two ways. Either a finger is briefly lifted and then replaced, or a finger that is currently lifted is used to briefly strike the hole. For some pitches only one or the other is possible, but for others both styles are possible. Being able to distinguish between the two from context may allow more detailed analysis of performance.

5 Conclusion

This paper has described a method for labelling gracenotes in musical performances. Since gracenotes are typically the shortest duration notes in a melody, clustering notes based on their duration allows the identification of the shortest notes. This is done using a mixture of distributions model and the EM algorithm. However, it was discovered that there may be an overlap between the component distributions representing the shortest melody notes and the longest gracenotes, making the labelling of some notes ambiguous. Since note duration alone is insufficient to do the labelling, a Bayesian method for incorporating contextual information into a posterior probability estimate is developed and found to be a substantial improvement. Finally, suggestions for extending the model to include a Markov random field and to handle multiple-gracenote clusters are presented.

6 Acknowledgements

The author would like to thank Brendan Frey for his thoughts on the form of the graphical model, and Robert MacNeil for supplying the database of tune scores. This work was supported in part by NSERC.

References

[1] Christopher Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1996.

[2] Roderick J. Cannon. *The Highland Bagpipe and its Music.* John Donald, 2000.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[4] Brendan J. Frey. *Graphical Models for Machine Learning and Digital Communication.* MIT Press, 1998.

[5] Michael I.Jordan, editor. *Learning in Graphical Models.* MIT Press, 1999.

[6] Geoffrey J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering.* Marcel Dekker Inc., New York, 1988.

[7] Christopher Raphael. A hybrid graphical model for rhythmic parsing. To appear in *Artificial Intelligence.*

[8] Bruno H. Repp. Further perceptual evaluations of pulse microstructure in computer performances of classical piano music. *Music Perception*, 8(1):1–33, 1990.

[9] Bruno H. Repp. Patterns of expressive timing in performances of a beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America*, 88(2):622–641, 1990.

[10] Bruno H. Repp. The detectability of local deviations from a typical expressive timing pattern. *Music Perception*, 15(3):265–289, 1998.

[11] Andrew Sterian and Gregory H. Wakefield. Music transcription systems: From sound to symbol. In *Proceedings of the AAAI-2000 Workshop on Artificial Intelligence and Music*, July 2000.

[12] C. Tait and William Findlay. Wavelet analysis for onset detection. In *Proceedings of the International Computer Music Conference, Hong Kong*, pages 500–503, 1996.

[13] Roland Wilson, Andrew D. Calway, and Edward S. Pearson. A generalized wavelet transform for fourier analysis: The multiresolution fourier transform and its application to image and audio signal analysis. *IEEE Transactions on Information Theory*, 38(2):674–690, 1992.