This is an electronic version of an article published in: Journal of New Music Research, 39(3), 203-213, 2010. doi:10.1080/09298215.2010.512979. Journal of New Music Research is available online at: www.tandfonline.com OpenURL: http://www.tandfonline.com/openurl?genre=article&issn=1744-5027&volume=39&issue=3&spage=203

Delayed decision-making in real-time beatbox percussion classification

Dan Stowell and Mark D. Plumbley

July 2010

Abstract

Real-time classification applied to a vocal percussion signal holds potential as an interface for live musical control. In this article we propose a novel approach to resolving the tension between the needs for low-latency reaction and reliable classification, by deferring the final classification decision until after a response has been initiated. We introduce a new dataset of annotated human beatbox recordings, and use it to study the optimal delay for classification accuracy. We then investigate the effect of such delayed decision-making on the quality of the audio output of a typical reactive system, via a MUSHRA-type listening test. Our results show that the effect depends on the output audio type: for popular dance/pop drum sounds the acceptable delay is on the order of 12–35 ms.

1 Introduction

In real-time signal processing it is often useful to identify and classify events represented within a signal. With music signals this need arises in applications such as live music transcription [Brossier, 2007] and human-machine musical interaction [Collins, 2006, Aucouturier and Pachet, 2006].

Yet to respond to events in real time presents a dilemma: often we wish a system to react with low latency, perhaps as soon as the beginning of an event is detected, but we also wish it to react with high precision, which may imply waiting until all information about the event has been received so as to make an optimal classification. The acceptable balance between these two demands will depend on the application context. In music, the perceptible event latency can be held to be around 30 ms, depending on the type of musical signal [Mäki-Patola and Hämäläinen, 2004].

We propose to deal with this dilemma by allowing event triggering and classification to occur at different times, thus allowing a fast reaction to be combined with an accurate classification. Triggering prior to classification implies that for a short period of time the system would need to respond using only a provisional classification, or some generic response. It could thus be used in reactive music systems if it were acceptable for some initial sound to be emitted even if the system's decision might change soon afterwards and the output updated accordingly. To evaluate such a technique applied to real-time music processing, we need to understand not only the scope for improved classification at increased latency, but

also the extent to which such delayed decision-making affects the listening experience, when reflected in the audio output.

In this paper we investigate delayed decision-making in the context of musical control by vocal percussion in the "human beatbox" style [Stowell, 2010, Section 2.2]. We consider the imitation of drum sounds commonly used in Western popular music such as kick (bass) drum, snare and hihat (for definitions of drum names see Randel [2003]). The classification of vocal sounds into such categories offers the potential for musical control by beatboxing, and some work has explored this potential in non-real-time [Sinyor et al., 2005] and in real-time [Hazan, 2005, Collins, 2004].

This paper investigates two aspects of the delayed decision-making concept. In Section 2 we study the relationship between latency and classification accuracy: we present an annotated dataset of human beatbox recordings, and describe classification experiments on these data. Then in Section 3 we describe a perceptual experiment using sampled drum sounds as could be controlled by live beatbox classification. The experiment investigates bounds on the tolerable latency of decision-making in such a context, and therefore the extent to which delayed decision-making can help resolve the tension between a system's speed of reaction and its accuracy of classification.

2 Classification experiment

We wish to be able to classify percussion events in an audio stream such as beatboxing, for example a three-way classification into kick/hihat/snare event types. We might apply an onset detector to detect events, then use acoustic features measured from the audio stream at the time of onset as input to a classifier which has been trained using appropriate example sounds [Hazan, 2005]. In such an application there are many options which will bear upon performance, including the choice of onset detector, acoustic features, classifier and training material. In the present work we factor out the influence of the onset detector by using manually-annotated onsets, and we introduce a real-world dataset for beatbox classification which we describe below.

We wish to investigate the hypothesis that the performance of some real-time classifier would improve if it were allowed to delay its decision so as to receive more information. In order that our results may be generalised we will use a classifier-independent measure of class separability, as well as results derived using a specific (although general-purpose) classifier.

To estimate class separability independent of a classifier we use the Kullback-Leibler divergence (KL divergence, also called the relative entropy) between the continuous feature distributions for classes [Cover and Thomas, 2006, section 9.5]:

$$D_{KL}(f||g) = \int f \log \frac{f}{g} \tag{1}$$

where f and g are the densities of the features for two classes. The KL divergence is an information-theoretic measure of the amount by which one probability distribution differs from another. It can be estimated from data with few assumptions about the underlying distributions, so has broad applicability. It is nonnegative and non-symmetric, although can be symmetrised by taking the value $D_{KL}(f||g) + D_{KL}(g||f)$ [Arndt, 2001, section 9.2]; in the present experiment we will further symmetrise over multiple classes by averaging D_{KL} over all class pairs to give a summary measure of the separability of the distribu-

tions. Because of the difficulties in estimating high-dimensional densities from data [Hastie et al., 2001, chapter 2] we will use divergence measures calculated for each feature separately, rather than in the high-dimensional joint feature space. Note that treating each feature separately will fail to detect some effects on separability caused by feature interactions. Such interaction effects rarely have a large impact, but would be worth studying in future.

To provide a more concrete study of classifier performance we will also apply a Naïve Bayes classifier [Langley et al., 1992], which estimates distributions separately for each input feature and then derives class probabilities for a datum simply by multiplying together the probabilities due to each feature. This classifier is selected for multiple reasons:

- It is a relatively simple and generic classifier, and well-studied, and so may be held to be a representative choice;
- Despite its simplicity and unrealistic assumptions (such as independence of features), it often achieves good classification results even in cases where its assumptions are not met [Domingos and Pazzani, 1997];
- The independence assumption makes possible an efficient updateable classifier in the real-time context: the class probabilities calculated using an initial set of features can be later updated with extra features, simply by multiplying by the probabilities derived from the new set of features.

Both our KL divergence estimates and our Naïve Bayes classification results operate on features independently. In this work we do not consider issues of redundancy between features.

2.1 Human beatbox dataset: *beatboxset1*

To facilitate the study of human beatbox audio we have collected and published a dataset which we call *beatboxset1*.[1] It consists of short recordings of beatboxing recorded by amateur and semi-professional beatboxers recorded under heterogenous conditions, as well as onset times and event classification annotations marked by independent annotators. The audio and metadata are freely available and published under the Creative Commons Attribution-Share Alike 3.0 license.

Audio: The audio files are 14 recordings each by a different beatboxer, between 12 and 95 seconds in length (mean duration 47 seconds). Audio files were recorded by the contributors, in a range of conditions: differing microphone type, recording equipment and background noise levels. The clips were provided by users of the website `humanbeatbox.com`.

Annotations: Annotations of the beatbox data were made by two independent annotators. Individual event onset locations were annotated, along with a category label. The labels used are given in Table 1. Files were annotated using Sonic Visualiser 1.5,[2] via a combination of listening and inspection of waveforms/spectrograms. A total of 7460 event annotations were recorded (3849 from one annotator, 3611 from the other).

The labelling scheme we propose in Table 1 was developed to group sounds into the main categories of sound heard in a beatboxing stream, and to provide for efficient data entry by annotators. For comparison, the table also lists the labels used for a five-way classification by Sinyor et al. [2005], as well as sym-

[1] `http://archive.org/details/beatboxset1`
[2] `http://sonicvisualiser.org`

Table 1: Event labelling scheme used in *beatboxset1*, and the frequencies of occurrence of each class label in the annotations.

Label	Description	SBN	Sinyor	Count
k	Kick	b / .	kick	1623
hc	Hihat, closed	t	closed	1840
ho	Hihat, open	tss	open	376
sb	Snare, *bish* or *pss*-like	psh	p-snare	469
sk	Snare, *k*-like (clap or rimshot snare sound)	k	k-snare	1025
s	Snare but not fitting the above types	–	–	181
t	Tom	–	–	201
br	Breath sound (not intended to sound like percussion)	h	–	132
m	Humming or similar (a note with no drum-like or speech-like nature)	m	–	404
v	Speech or singing	[words]	–	76
x	Miscellaneous other sound	–	–	1072
?	Unsure of classification	–	–	61

bols from Standard Beatbox Notation (SBN – a simplified type of score notation for beatbox performers).[3] Our labelling is oriented around the sounds produced rather than the mechanics of production (as in SBN), but aggregates over the fine phonetic details of each realisation (as would be shown in an International Phonetic Alphabet transcription).

The final column in Table 1 gives the frequency of occurrence of each of the class labels, confirming that the majority (74%) of the events fall broadly into the kick, hihat, and snare categories.

2.2 Method

To perform a three-way classification experiment on *beatboxset1* we aggregated the labelled classes into the three main types of percussion sound:

- kick (label `k`; 1623 instances),
- snare (labels `s`, `sb`, `sk`; 1675 instances),
- hihat (labels `hc`, `ho`; 2216 instances).

The events labelled with other classes were not included in the present experiment.

[3] `http://www.humanbeatbox.com/tips/`

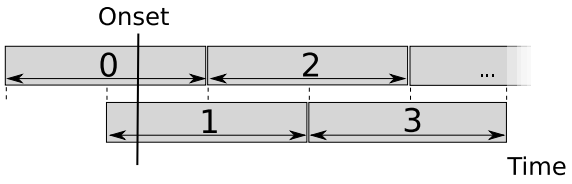


Figure 1: Numbering the "delay" of audio frames relative to the temporal location of an annotated onset.

We analysed the soundfiles to produce the set of 24 features listed in Table 2. Features were derived using a 44.1 kHz audio sampling rate, and a frame size of 1024 samples (23 ms) with 50% overlap (giving a feature sampling rate of 86.1 Hz).

Each manually-annotated onset was aligned with the first audio frame containing it (the earliest frame in which an onset could be expected to be detected in a real-time system). In the following, the amount of delay will be specified in numbers of frames relative to that aligned frame, as illustrated in Figure 1. We investigated delays of zero through to seven frames, corresponding to a latency of 0–81 ms.

In applying the Naïve Bayes classifier, we investigated four different strategies for choosing features as

Table 2: Acoustic features measured (for definitions of many of these see Peeters [2004]; *HFC* and *flux* are as in [Brossier, 2007, section 2.3], crest features are as in [Hosseinzadeh and Krishnan, 2008])

Label	Feature
mfcc1–mfcc8	Eight MFCCs, derived from 42 Mel-spaced filters (zero'th MFCC not included)
centroid	Spectral centroid
spread	Spectral spread
scf	Spectral crest factor
scf1–scf4	Spectral crest factor in subbands (50–400, 400–800, 800–1600, and 1600–3200 Hz)
25%ile–95%ile	Spectral distribution percentiles: 25%, 50%, 90%, 95% ("rolloff")
HFC	High-frequency content
ZCR	Zero-crossing rate
flatness	Spectral flatness
flux	Spectral flux
slope	Spectral slope

input to the classifier – with or without stacking, and with two types of feature selection:

Feature stacking: We first used only the features derived from the frame at a single delay value (as with the divergence measures above). However, as we delay the decision the information from earlier frames is in principle available to the classifier, so we should be able to improve classification performance by making use of this extra information – in the simplest case by "stacking" feature values, creating a larger featureset from the concatenation of the features from multiple frames [Meng, 2006, Section 4.2]. (This is termed "shingling" by Casey et al. [2008].) Therefore we also performed classification at each delay using the fully stacked featuresets, aggregating all frames from onset up to the specified delay. Our 24-feature set at zero delay would become a 48-feature set at one frame delay, then a 72-feature set at two frames' delay, and so forth.

Feature selection: Stacking features creates very large featuresets and so risks incurring curse-of-dimensionality issues, well known in machine learning: large dimensionalities can reduce the effectiveness of classifiers, or at least require exponentially more training data to prevent overfitting [Hastie et al., 2001, chapter 2]. To circumvent the curse of dimensionality yet combine information from different frames, we applied two forms of feature-selection. The first used each of our 24 features once only, but taken at the amount of delay corresponding to the best class separability for that feature. The second applied a standard feature-selection algorithm to choose the 24 best features at different delays, allowing it to choose a feature multiple times at different delays. We used the Information Gain selection algorithm [Mitchell, 1997, section 3.4.1] for this purpose.

To estimate the KL divergence from data, we used a Gaussian kernel estimate for the distribution of each feature separately for each class. For each feature we then estimated the KL divergence pairwise between classes, by numerical integration over the estimated distributions (since the KL divergence is a directed measure, there are six pairwise measures for the three classes). To summarise the separability of the three classes we report the mean of the six estimated divergences, which gives a symmetrised measure of divergence between the three classes. Since our KL divergence measure treats each single feature independently, stacking and feature-selection are not relevant and were not applied.

Implementation: We used SuperCollider 3.3 [McCartney, 2002] for feature analysis, with Hann windowing applied to frames before spectral analysis. KL divergence was estimated using *gaussian_kde* from the SciPy 0.7.1 package, running in

Python 2.5.4, with bandwidth selection by Scott's Rule. Classification experiments were performed using Weka 3.5.6 [Witten and Frank, 2005], using ten-fold cross-validation.

2.3 Results

The class separability measured by average KL divergence between classes is given in Figure 2, and the peak values for each feature in Table 3. The values of the divergences cover a broad range depending on both the feature type and the amount of delay, and in general a delay of around 2 frames (23 ms) appears under this measure to give the best class separation. Note that this analysis considers each amount of delay separately, ignoring the information available in earlier frames. The separability at zero delay is generally the poorest of all the delays studied here, which is perhaps unsurprising, as the audio frame containing the onset will often contain a small amount of unrelated audio prior to the onset plus some of the quietest sound in the beginning of the attack. The peak separability for the features appears to show some variation, occurring at delays ranging from 1 to 4 frames. The highest peaks occur in the spectral 25- and 50-percentile (at 3 frames' delay), suggesting that the distribution of energy in the lower part of the spectrum may be the clearest differentiator between the classes.

Table 3: The delay giving the peak symmetrised KL divergence for each feature.

Feature	Delay	Divergence
mfcc1	3	1.338
mfcc2	3	0.7369
mfcc3	1	0.3837
mfcc4	3	0.1747
mfcc5	1	0.2613
mfcc6	6	0.2512
mfcc7	1	0.1778
mfcc8	2	0.312
centroid	3	1.9857
spread	2	0.5546
scf	2	0.6975
scf1	0	0.1312
scf2	2	0.0658
scf3	4	0.0547
scf4	4	0.0929
25%ile	3	4.6005
50%ile	3	2.9217
90%ile	2	0.8857
95%ile	2	0.6427
HFC	4	0.7245
ZCR	1	0.454
flatness	2	0.6412
flux	1	1.2058
slope	1	1.453

The class separability measurements are reflected in the performance of the Naïve Bayes classifier on our three-way classification test (Figure 3). When using only the information from the latest frame at each delay the data show a similar curve: poor performance at zero delay, rising to a strong performance at 1 to 3 frames' delay (peaking at 75.0% for 2 frames), then tailing off gradually at larger delays.

When using feature stacking the classifier is able to perform strongly at the later delays, having access to information from the informative early frames, although a slight curse-of-dimensionality effect is visible in the very longest delays we investigated: the classification accuracy peaks at 5 frames (77.6%) and tails off afterwards, even though the classifier is given the exact same information plus some extra features. Overall, the improvement due to feature stacking is small compared against the single-frame peak performance. Such a small advantage would need to be balanced against the increased memory requirements and complexity of a classifier implemented in a real-time system – although as previously mentioned,

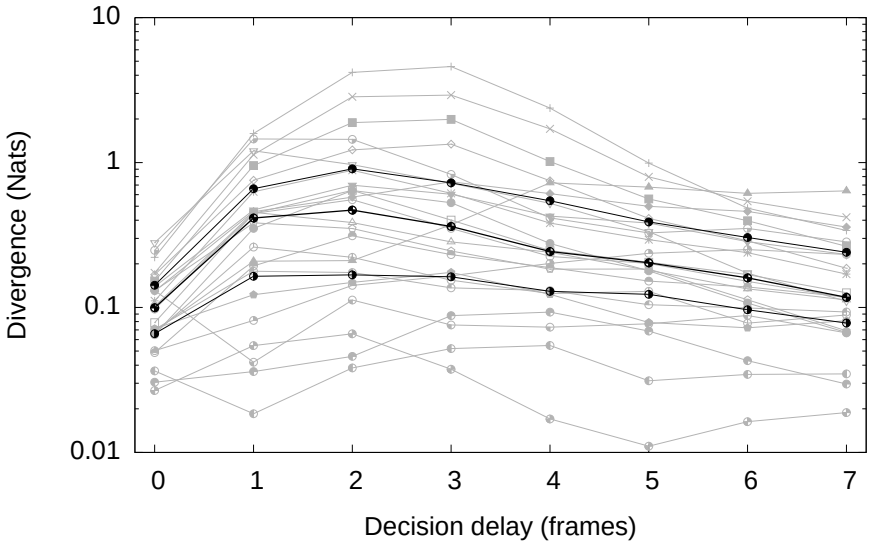


Figure 2: Separability measured by average KL divergence, as a function of the delay after onset. At each frame the class separability is summarised using the feature values measured only in that frame. The grey lines indicate the individual divergence statistics for each of the 24 features, while the dark lines summarise over all features, showing the median and the 25- and 75-percentiles of the symmetrised divergence measure.

the independence assumption of the classifier allows frame information to be combined at relatively low complexity.

We also performed feature selection as described earlier, first using the peak-performing delays given in Table 3 and then using features/delays selected using Information Gain (Table 4). In both cases some of the selected features are unavailable in the earlier stages so the feature set is of low dimensionality, only reaching 24 dimensions at the 5- or 6-frame delay point. The performance of these sets shows a similar trajectory to the full stacked feature set although consistently slightly inferior to it. The Information Gain approach is in a sense less constrained than the former approach – it may select a feature more than once at different delays – yet does not show superior performance, suggesting that the variety of features is more important than the varieties of delay in classification performance.

The Information Gain feature selections (Table 4) also suggest which of our features may be generally best for the beatbox classification task. The 25- and 50-percentile are highly ranked (confirming our observation made on the divergence measures), as are the spectral centroid and spectral flux.

A confusion matrix for the classifier output at the peak-performing delay of 2 frames (for the non-stacked feature set) is given in Table 5, revealing a particular tendency for snare sounds to be misclassified as kick sounds. To probe the differences in separability between different class pairs, as a follow-up we investigated the performance of the classifier on each of the two-class sub-tasks (hihat vs. others, kick vs. others, snare vs. others). The results (Figure

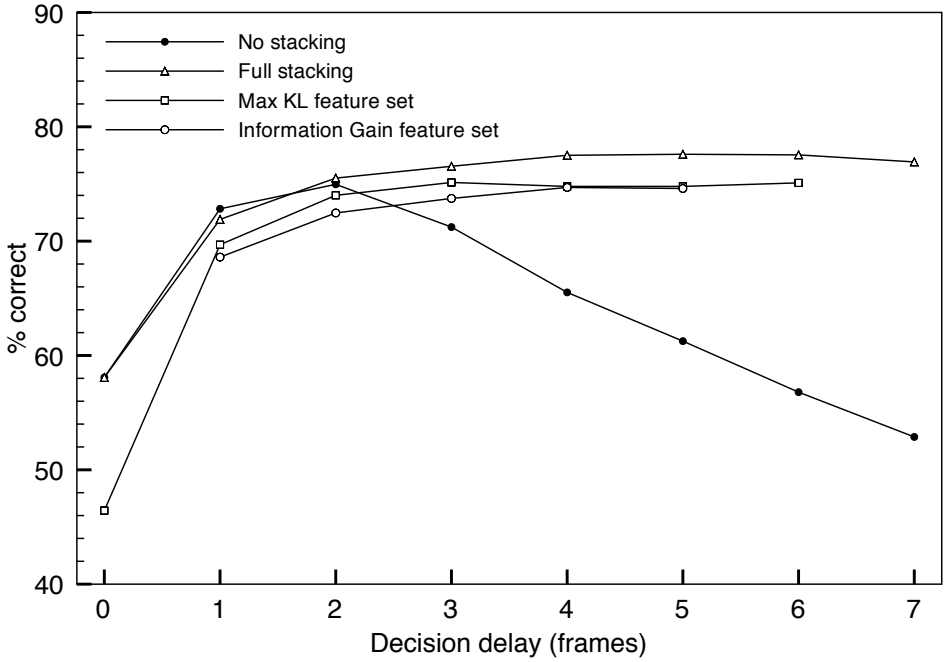


Figure 3: Classification accuracy using Naïve Bayes classifier (3 classes)

4, upper) show a clear difference between the subtasks: the classifier is able to distinguish the hihat from either of the other two classes with a high degree of success at 1 or 2 frames delay, while the classification of kicks peaks at around 2–3 frames, and of snares around 4 frames. The snare vs. others subtask shows bimodal results. When we plot the performance of the two-class sub-tasks created by excluding one class of events entirely (Figure 4, lower), we see the bimodality seems due to the strong hihat/snare distinction which can be made as early as 1 frame with the kick/snare distinction peaking much later (4 frames, ~ 50 ms) and at a lower accuracy.

These results suggest either that the attack segments of kick and snare beatboxing sounds are broadly similar to each other and different from those of hihat sounds, and the differences emerge mainly during the decay segment; or that there are differences which are not captured by our feature set. We suggest the former may be the dominating factor, because both kick and snare sounds can be produced with bilabial plosive onsets (*k* and *sb* in Table 1). Others have studied classification of non-beatbox drum sounds based on brief attack segments, with acceptable results (depending on the exact task) [Tindale et al., 2004, Pachet and Roy, 2009]. Beatboxing may be a more challenging classification task than other percussion because all sounds are produced by the same apparatus in various configurations, rather than by different sounding bodies.

In summary, we find that with this dataset of beatboxing recorded under heterogeneous conditions, a delay of around 2 frames (23 ms) relative to onset leads to stronger performance in a three-way classi-

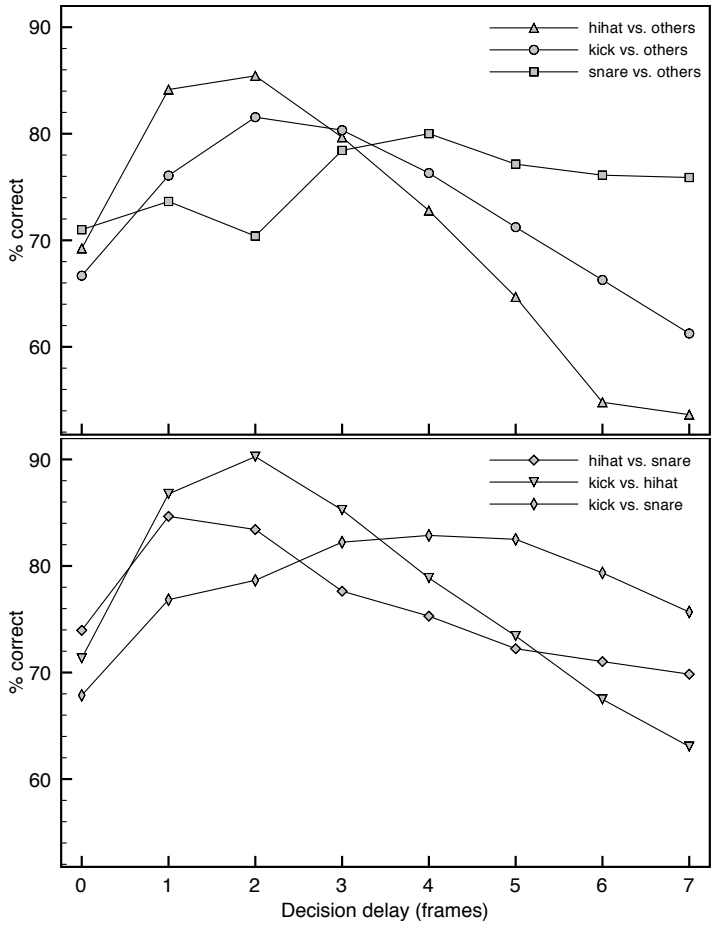


Figure 4: Classification accuracy using Naïve Bayes classifier on two-class sub-tasks (all features, no stacking)

fication task. (Compare e.g. Brossier [2007, section 5.3.3], who finds that for real-time pitch-tracking of musical instruments, reliable note estimation is not possible until around 45 ms after onset.) Feature stacking further improves classification results for decisions delayed by 2 frames or more, although at the cost of increased dimensionality of the feature space. Reducing the dimensionality by feature selection over the different amounts of delay can provide good classification results at large delays with low complexity, but fails to show improvement over the classifier performance simply using the features at the best delay of 2 frames.

In designing a system for real-time beatbox classification, then, a classification at the earliest possible opportunity is likely to be suboptimal, especially when using known onsets or an onset detector designed for low-latency response. Classification delayed until roughly 10–20 ms after onset detection would provide better performance. Features characterising the distribution of the lower-frequency energy (the spectral 25- and 50-percentiles and centroid) can

Table 4: The 24 features and delays selected using Information Gain, out of a possible 192.

Rank	Feature	Delay	Rank	Feature	Delay
1	*50%ile*	2	13	*mfcc1*	2
2	*centroid*	2	14	*90%ile*	2
3	*50%ile*	3	15	*slope*	2
4	*centroid*	3	16	*25%ile*	1
5	*25%ile*	2	17	*50%ile*	5
6	*flux*	1	18	*flux*	3
7	*flux*	2	19	*ZCR*	1
8	*50%ile*	4	20	*25%ile*	4
9	*50%ile*	1	21	*centroid*	4
10	*slope*	1	22	*mfcc1*	1
11	*centroid*	1	23	*mfcc1*	3
12	*25%ile*	3	24	*90%ile*	1

Table 5: Confusion matrix for the Naïve Bayes classifier at 2 frames delay and with no stacking. Rows indicate the ground-truth label, and columns the classifier decision.

%	*kick*	*hihat*	*snare*
kick	88.4	1.9	8.8
hihat	16.1	81.6	8.7
snare	31.5	12.4	53.1

be recommended for this task.

Since the Naïve Bayes classifier treats features independently, a real-time system could progressively update the classification decision as each new frame arrives, progressively increasing the amount of stacking. In fact, the two-way classification results indicate that the classification task could be spread across frames, using a decision-tree approach [Murthy, 1998] in which a hihat-vs.-others decision could be made at a low latency, and the snare-vs.-kick decision made slightly later. In Section 3 we will study the perceptual quality of a system whose decision is only updated once, in order to create a clear experimental measure of the relationship between delay and quality. However we note that a progressively-updated decision is a useful possibility for the real-time classification task discussed here, to be explored in future work.

3 Perceptual experiment

In Section 2 we confirmed that beatbox classification can be improved by delaying decision-making relative to the event onset. Adding this extra latency to the audio output may be undesirable in a real-time percussive performance, hence our proposal that a low-latency low-accuracy output could be updated some milliseconds later with an improved classification. This two-step approach would affect the nature of the output audio, so we next investigate the likely effect on audio quality via a listening test.

Our test will be based on the model of a reactive musical system which can trigger sound samples, yet which allows that the decision about which sound sample to trigger may be updated some milliseconds later. Between the initial trigger and the final classification the system might begin to output the most likely sample according to initial information, or a mixture of all the possible samples, or some generic "placeholder" sound such as pink noise. The resulting audio output may therefore contain some degree of inappropriate or distracting content in the attack segments of events. It is known that the attack portion of musical sounds carries salient timbre information, although that information is to some extent redundantly distributed across the attack and later portions of the sound [Iverson and Krumhansl, 1993]. Our research question here is the extent to which the inappropriate attack content introduced by delayed decision-making impedes the perceived quality of the audio stream produced.

3.1 Method

We first created a set of audio stimuli for use in the listening test. The delayed-classification concept was implemented in the generation of a set of drum loop recordings as follows: for a given drum hit, the desired sound (e.g. kick) was not output at first, but rather an equal mixture of kick, hihat and snare sounds. Then after the chosen delay time the mixture was crossfaded (with a 1ms sinusoidal crossfade) to become purely the desired sound. The resulting signal could be considered to be a drum loop in which the onset timings were preserved, but the onsets of the samples had been degraded by contamination with other sound samples. We investigated amounts of delay corresponding to 1, 2, 3 and 4 frames as in the earlier classifier experiment (Section 2) - approximately 12, 23, 35 and 46 ms.

Sound excerpts generated by this method therefore represent a kind of idealised and simplified delayed decision-making in which no information is available at the moment of onset (hence the equal balance of all drum types) and 100% classification accuracy occurs after the specified delay. Our classifier experiment (Section 2) indicates that in a real-time classification system, some information is available soon after onset, and also that classification is unlikely to achieve perfect classification accuracy. The current experiment factors out such issues of classifier performance to focus on the perceptual effect of delayed decision-making in itself.

The reference signals were each 8 seconds of drum loops at 120bpm with one drum sample (kick/snare/hihat) being played on every eighth-note. Three drum patterns were created using standard dance/pop rhythms, such that the three classes of sound were equally represented across the patterns. The patterns were (using notation k=kick, h=hihat, s=snare):

```
k k s h h k s h
k h s s k k s h
k h s k h s h s
```

We created the sound excerpts separately with three different sets of drum sound samples, which were chosen to be representative of standard dance/pop drum sounds as well as providing different levels of susceptibility to degradation induced by delayed classification:

Immediate-onset samples, designed by the first author using SuperCollider to give kick/hihat/snare sounds, but with short duration and zero attack time, so as to provide a strong test for the delayed classification. This drum set was expected to provide poor acceptability at even moderate amounts of delay.

Roland TR909 samples, taken from one of the most popular drum synthesisers in dance music [Butler, 2006, p. 326], with a moderately realistic sound. This drum set was expected to provide moderate acceptability results.

Amen break, originally sampled from "Amen brother" by The Winstons and later the basis of jungle, breakcore and other genres, now the most popular breakbeat in dance music [Butler, 2006, p. 78]. The sound samples are much less "clean" than the other sound samples (all three samples clearly contain the sound of a ride cymbal, for example). Therefore this set was expected to provide more robust acceptance results than the other sets, yet still represent a commonly-used class of drum sound.

The amplitude of the three sets of audio excerpts was adjusted manually by the first author for equal loudness.

Tests were performed within the "MUlti Stimulus test with Hidden Reference and Anchor" (MUSHRA) standard framework [International Telecommunication Union, 2003]. In the MUSHRA test participants are presented with sets of processed audio excerpts and asked to rate their *basic audio quality* in relation to a reference unprocessed audio excerpt. Each set of excerpts includes the unprocessed audio as a hidden reference, plus a 3.5 kHz low-pass filtered version of the excerpt as a low-quality anchor, as well as excerpts produced by the systems investigated.

Our MUSHRA tests were fully balanced over all combinations of the three drum sets and the three patterns, giving nine trials in total. In each trial, participants were presented with the unprocessed reference excerpt, plus six excerpts to be graded: the hidden reference, the filtered anchor, and the delayed-decision versions at 1, 2, 3 and 4 frames' delay (see Figure 5 for a screenshot of one trial). The order of the trials and of the excerpts within each trial was randomised.

Participants: We recruited 23 experienced music listeners (17 men and 6 women) aged between 23 and 43 (mean age 31.3). Some participants had experience as musicians; none were beatboxers, and a minority (two) had experience playing percussion. Tests took around 20–30 minutes in total to complete, including initial training, and were performed using headphones.

Post-screening was performed by numerical tests combined with manual inspection. For each participant we calculated correlations (Pearson's r and Spearman's ρ) of their gradings with the median of the gradings provided by the other participants. Any

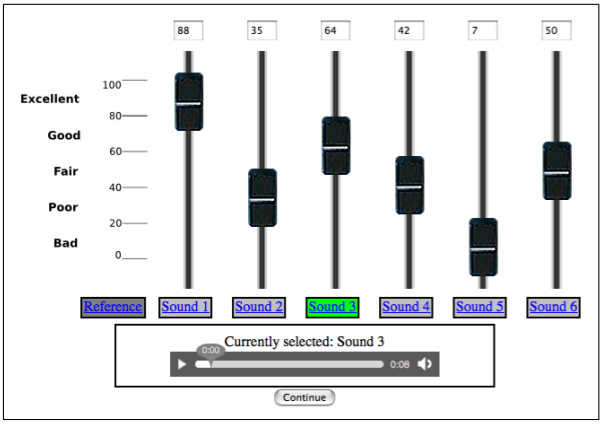


Figure 5: The user interface for one trial within the MUSHRA listening test.

set of gradings with a low correlation was inspected as a possible outlier. Any set of gradings in which the hidden reference was not always rated at 100 was also inspected manually. (Ideally the hidden reference should always be rated at 100 since it is identical to the reference; however, participants tend to treat MUSHRA-type tasks to some extent as ranking tasks [Sporer et al., 2009], and so if they misidentify some other signal as the highest quality they may penalise the hidden reference slightly. Hence we did not automatically reject these.)

We also plotted the pairwise correlations between gradings for every pair of participants, to check for subgroup effects. No subgroups were found, and one outlier was identified. The remaining 22 participants' gradings were analysed as a single group.

The MUSHRA standard [International Telecommunication Union, 2003] recommends calculating the mean and confidence interval for listening test data. However, the grading scale is bounded (between 0 and 100) which can lead to difficulties using the standard normality assumption to calculate confidence intervals, especially at the extremes of the scale. To

mitigate these issues we applied the logistic transformation [Siegel, 1988, chapter 9]:

$$z = \log \frac{x + \delta}{100 + \delta - x}\,, \tag{2}$$

where x is the original MUSHRA score and the δ is added to prevent boundary values from mapping to $\pm\infty$ (we used $\delta = 0.5$). Such transformation allows standard parametric tests to be applied more meaningfully (see also Lesaffre et al. [2007]). We calculated our statistics (mean, confidence intervals, t-tests) on the transformed values z before projecting back to the original domain.

The audio excerpts, participant responses, and analysis script for this experiment are published online.[4]

3.2 Results

For each kit, we investigated the differences pairwise between each of the six conditions (the four delay levels plus the reference and anchor). To determine whether the differences between conditions were significant we performed the paired samples t-test (in the logistic z domain; d.f. = 65) with a significance threshold of 0.01, applying Holm's procedure to control for multiple comparisons [Shaffer, 1995]. All differences were found to be significant with the exception of the following pairs:

- Immediate-onset samples:
  - anchor and 12 ms
  - 23 ms and 35 ms
  - 35 ms and 46 ms
- Roland TR909 samples:
  - anchor and 35 ms
  - anchor and 46 ms

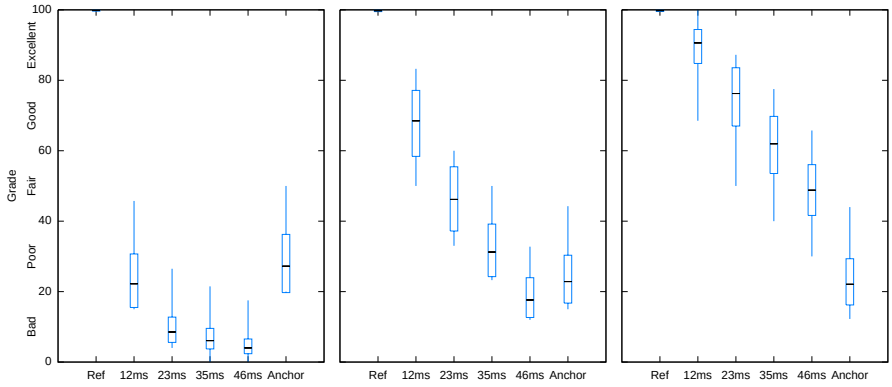
The logistic transformation mitigates against boundary effects when applying parametric tests. However the MUSHRA standard does not propose such transformation, so as an additional validation check we also applied the above test on the data in its original domain. In this instance the significance testing produced the same results.

Figure 6 summarises the results of the listening test. It confirms that for each of the drum sets, the degradation is perceptible by listeners since the reference is readily identifiable, and also that the listening quality becomes worse as the delay lengthens. It also demonstrates that the three drum sets vary in their robustness to this degradation, as expected.

The immediate-onset drum set was designed to provide a kind of lower bound on the acceptability, and it does indeed show very poor gradings under all of the delay lengths we investigated. Participants mostly found the audio quality to be worse than the low-pass filtered anchor, except in the 12 ms condition where no significant difference from the anchor was found, so we say that participants found the audio quality to be similarly poor as the anchor. For such a drum set, this indicates that delayed decision-making would likely be untenable.

The other two sets of drum sounds are more typical of drum sounds used in popular music, and both are relatively more robust to the degradation. Sound quality was rated as 60 or better (corresponding in the MUSHRA quality scale to *good* or *excellent*) at 12 ms for the TR909 set, and up as far as 35 ms for the Amen set. Even at 46 ms delay, the acceptability for the Amen set is much greater than that for the immediate-onset set at 12 ms delay.

[4] `http://archive.org/details/dsmushradata09`



(a) Immediate-onset (b) TR909 (c) Amen break

Figure 6: Results from the listening test, showing the mean and 95% confidence intervals (calculated in the logistic transformation domain) with whiskers extending to the 25- and 75-percentiles. The plots show results for the three drum sets separately. The durations given on the horizontal axis indicate the delay, corresponding to 1/2/3/4 audio frames in the classification experiment.

When applied in a real-world implementation, the extent to which these perceptual quality measures reflect the amount of delay acceptable will depend on the application. For a live performance in which real-time controlled percussion is one component of a complete musical performance, the delays corresponding to good or excellent audio quality could well be acceptable, in return for an improved classification accuracy without added latency.

4 Conclusions

We have investigated delayed decision-making in real-time classification, as a strategy to allow for improved characterisation of events in real-time without increasing the triggering latency of a system. This possibility depends on the notion that small signal degradations introduced by using an indeterminate onset sound might be acceptable in terms of perceptual audio quality.

We introduced a new real-world beatboxing dataset *beatboxset1* and used it to investigate the improvement in classification that might result from delayed decision-making on such signals. A delay of 23 ms generally performed strongly out of those we tested. Neither feature stacking nor feature selection across varying amounts of delay led to strong improvements over this performance, though some of the classification sub-tasks (hihat vs. others) showed peak performance at a lower delay compared to others (kick vs. snare), suggesting that the acoustic signal properties of the classes separate out at different stages.

In a MUSHRA-type listening test we then investigated the effect on perceptual audio quality of a degradation representative of delayed decision-making. We found that the resulting audio quality depended strongly on the type of percussion sound in use. The effect of delayed decision-making was

readily perceptible in our listening test, and for some types of sound delayed decision-making led to unacceptable degradation (poor/bad quality) at any delay; but for common dance/pop drum sounds, the maximum delay which preserved an excellent or good audio quality varied from 12 ms to 35 ms.

Acknowledgments

We thank the beatboxers featured in *beatboxset1* and the annotators Helena du Toit and Diako Rasoul, who were supported by a bursary from the Nuffield Foundation. We also thank an anonymous reviewer for suggesting the two-class analysis and its potential use in a progressive decision tree classifier. DS is supported by the EPSRC under a Doctoral Training Account studentship. MP is supported by an EPSRC Leadership Fellowship (EP/G007144/1).

References

C. Arndt. *Information Measures*. Springer, 2001.

J.-J. Aucouturier and F. Pachet. Jamming with plunderphonics: interactive concatenative synthesis of music. *Journal of New Music Research*, 35(1):35–50, Mar 2006.

P. M. Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Dept of Electronic Engineering, Queen Mary University of London, London, UK, Mar 2007. URL `http://aubio.piem.org/phdthesis/`.

M. J. Butler. *Unlocking the Groove: Rhythm, Meter, and Musical Design in Electronic Dance Music*. Indiana University Press, Bloomington, 2006.

M. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):1015–1028, Jul 2008. doi: 10.1109/TASL.2008.925883.

N. Collins. On onsets on-the-fly: real-time event segmentation and categorisation as a compositional effect. In *Proceedings of Sound and Music Computing*, pages 20–22, Oct 2004.

N. Collins. *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*. PhD thesis, University of Cambridge, 2006. URL `http://www.cogs.susx.ac.uk/users/nc81/thesis.html`.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience New York, 2nd edition, 2006.

P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2–3):103–130, 1997. doi: 10.1023/A:1007413511361.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2001.

A. Hazan. Billaboop: real-time voice-driven drum generator. In *Proceedings of the 118th Audio Engineering Society Convention (AES 118)*, number 6411, May 2005.

D. Hosseinzadeh and S. Krishnan. On the use of complementary spectral features for speaker recognition. *EURASIP Journal on Advances in Signal Processing*, 2008 (Article ID 258184, 10 pages), 2008. doi: 10.1155/2008/258184.

International Telecommunication Union. Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA). Technical Report ITU-R BS.1534-1, International Telecommunication Union, 2003. URL `http://www.itu.int/rec/R-REC-BS.1534/en`.

P. Iverson and C. L. Krumhansl. Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, 94(5):2595–2603, 1993. doi: 10.1121/1.407371.

P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 223–228, 1992.

E. Lesaffre, D. Rizopoulos, and R. Tsonaka. The logistic transform for bounded outcome scores. *Biostatistics*, 8(1):72–85, 2007. doi: 10.1093/biostatistics/kxj034.

T. Mäki-Patola and P. Hämäläinen. Latency tolerance for gesture controlled continuous sound instrument without tactile feedback. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 1–5, 2004.

J. McCartney. Rethinking the computer music language: SuperCollider. *Computer Music Journal*, 26(4):61–68, 2002. doi: 10.1162/014892602320991383.

A. Meng. *Temporal Feature Integration for Music Organisation*. PhD thesis, Technical University of Denmark (DTU), 2006. URL `http://www2.imm.dtu.dk/pubdb/p.php?4502`.

T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

S. K. Murthy. Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998. doi: 10.1023/A:1009744630224.

F. Pachet and P. Roy. Analytical features: a knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009(153017), 2009. doi: 10.1155/2009/153017.

G. Peeters. A large set of audio features for sound description. Technical report, IRCAM, 2004.

D. M. Randel. Drum set. In D. M. Randel, editor, *The Harvard Dictionary of Music*, page 256. Harvard University Press, 4th edition, 2003.

J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995. doi: 10.1146/annurev.ps.46.020195.003021.

A. F. Siegel. *Statistics and Data Analysis: An Introduction*. John Wiley & Sons Inc, New York, 1988.

E. Sinyor, C. McKay, R. Fiebrink, D. McEnnis, and I. Fujinaga. Beatbox classification using ACE. In *Proceedings of the International Conference on Music Information Retrieval*, pages 672–675, 2005.

T. Sporer, J. Liebetrau, and S. Schneider. Statistics of MUSHRA revisited. In *Proceedings of the 127th Audio Engineering Society Convention (AES 127)*, number 7825. Audio Engineering Society, Oct 2009.

D. Stowell. *Musical expression through real-time voice timbre analysis: machine learning and timbral control*. PhD thesis, School of Electronic Engineering and Computer Science, Queen Mary University

of London, 2010. URL `http://www.mcld.co.uk/thesis/`.

A. Tindale, A. Kapur, G. Tzanetakis, and I. Fujinaga. Retrieval of percussion gestures using timbre classification techniques. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, pages 541–545, 2004.

I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Technique*. Morgan Kaufmann, San Francisco, CA, USA, 2nd edition, 2005.