CAUSAL PREDICTION OF CONTINUOUS-VALUED MUSIC FEATURES

Peter Foster, Anssi Klapuri, Mark D. Plumbley
Centre for Digital Music
Queen Mary University of London
Mile End Road, London E1, UK
{peter.foster,anssi.klapuri,mark.plumbley}@eecs.qmul.ac.uk

ABSTRACT

This paper investigates techniques for predicting sequences of continuous-valued feature vectors extracted from musical audio. In particular, we consider prediction of beat-synchronous Mel-frequency cepstral coefficients and chroma features in a causal setting, where features are predicted as they unfold in time. The methods studied comprise autoregressive models, N-gram models incorporating a smoothing scheme, and a novel technique based on repetition detection using a self-distance matrix. Furthermore, we propose a method for combining predictors, which relies on a running estimate of the error variance of the predictors to inform a linear weighting of the predictor outputs. Results indicate that incorporating information on long-term structure improves the prediction performance for continuous-valued, sequential musical data. For the Beatles data set, combining the proposed self-distance based predictor with both N-gram and autoregressive methods results in an average of 13% improvement compared to a linear predictive baseline.

1. INTRODUCTION

Our goal is to devise methods for predicting music in a causal setting. Given a stream of observed music feature vectors extracted from an audio signal, we seek to predict future values of feature vectors. Furthermore, we seek to incorporate domain knowledge about the underlying music signal into the prediction process: Across musical genres, music exhibits hierarchical temporal structure, arising centrally from the identity relations between structural elements [11]. In Western music, elementary events are typically rhythmic, melodic or harmonic and give rise to long-term structure characteristic of a piece's musical form, through application of variation and repetition. Conversely, identifying parallelism in music — the occurrence of variation and repetition — is agreed to bear great importance in music-theoretical analysis [2].

This view may be considered to encompass cognitive processes involved in music listening. Here, music consists of a stream of events unfolding in time and experienced by a listener [9]. The listening process is associated with predictions of future events, which depend on the listener's evolving internal model of musical structure generated by previously observed events in the stream of music. This work is based on this causal prediction setting, where at a given point in time only events in the past inform predictions.

Accurate prediction of spectro-temporal features, such as Mel-frequency cepstral coefficients (MFCCs), chroma or rhythmograms [15], is motivated by a number of applications. Firstly, audio visualisation tasks might benefit from prediction, since live performance environments typically constrain the permissible amount of latency introduced in the audio processing chain [6]. Similarly, it is of interest to investigate robust real-time audio streaming applications for live music performance [10]. In the latter case, employing prediction techniques might allow the effect of network latency to be offset. Further applications of audio based prediction are automated musical accompaniment [8, 20] and audio feature models for automated music transcription.

In addition, prediction accuracy can be related to the assumed model of the underlying distribution of observations. In terms of inductive inference [19], accurate prediction relates to effective data compression of observations. This relationship might be exploited in online music content analysis applications. Existing work has examined the problem of offline music content analysis, where compressibility is used to evaluate structural similarity between pieces of music [1]. A related application is information-dynamic modelling of musical audio [4].

In this work, we evaluate several prediction methods, including autoregressive models, N-gram models, and a novel technique based on utilising the long-term structure of music signals. In addition, we propose a method for combining predictors by estimating predictors' error variance. We consider chroma and MFCC features, which describe harmonic and timbral information in musical audio signals [15]. Re-

sults indicate that combining the self-distance approach with autoregressive or N-gram models substantially improves the accuracy of predicting continuous-valued music features.

1.1 Causal music feature prediction

Suppose we have a sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_T$, corresponding to T feature observations made at times $\tau_1, \ldots, \tau_T$. Each vector occupies k-dimensional feature space, $\mathbf{v} \in \mathbb{R}^k$, according to an unknown probability distribution. Causal prediction involves approximating the unknown conditional probability distribution $p(\mathbf{v}_t \,|\, \mathbf{v}_1, \ldots, \mathbf{v}_{t-1})$. The predicted feature at time τ_t is then obtained by computing the expectation $\mathbf{E}[\mathbf{v}_t|\mathbf{v}_1, \ldots, \mathbf{v}_{t-1}]$. The prediction task is causal, since observations $\mathbf{v}_1, \ldots, \mathbf{v}_{t-1}$ inform predictions $\mathbf{v}_t$. Successive predictions are formed by increasing t, so that the observation history accumulates over time.

Causal predictive models have been applied to music in symbolic formats [16]. In the audio domain, the concern of our presented work, [8] proposes an approach for prediction driven musical expectation modelling. In [3] prediction is examined in the context of planning, as a means of creating anticipatory music systems. In [20] a method is proposed for automatic harmonic accompaniment based on repetition detection.

2. PREDICTION TECHNIQUES

We investigate prediction techniques for beat-synchronous chroma and MFCC features, as described in the following.

2.1 Autoregressive models

In a multivariate autoregressive (MAR) model [12], predicted feature vectors $\mathbf{v}_t$ are computed as linear combinations of N preceding feature vectors' components. Correlation between separate components is taken into account, so that

$$\mathbf{v}_t = \sum_{n=1}^{N} \mathbf{A}_n \mathbf{v}_{t-n} + \mathbf{r}_t \tag{1}$$

where matrices $\mathbf{A}_n$ incorporate information on correlations between between components of $\mathbf{v}_{t-n}$ and $\mathbf{v}_t$. Vector $\mathbf{r}_t$ is an independent and identically distributed Gaussian noise term.

Let us use $v_{t,u}$ to denote the uth component of vector $\mathbf{v}_t$. A special case of the MAR model arises when independence between feature components is assumed. In that case, matrices $\mathbf{A}_n$ are diagonal, so that

$$v_{t,u} = \sum_{n=1}^{N} a_{n,u} v_{t-n,u} + r_{t,u} \tag{2}$$

with $1 \leq u \leq k$. Coefficients $r_{t,u}$ are described by k univariate Gaussian noise processes with finite mean and variance. The model in Equation 2 is equivalent to a component-wise linear predictive coding (LPC) model, with each LPC model defined by index u.

2.2 N-gram prediction

N-gram models have been used to model symbolic music [16]. In this model, observations are quantised. Let e_t denote a quantised observation symbol. Symbols are members of a specified alphabet $\mathcal{A}$. For convenience, we use e_{t-n}^{t-1} to denote the sequence of symbols $e_{t-n}, e_{t-n+1}, \ldots, e_{t-1}$. The conditional probability of predicted event e_t, given the history of observations is assumed to obey the Markov property. That is, $p(e_t|e_1^{t-1}) = p(e_t|e_{t-n}^{t-1})$, where n is the order of the Markov model. An estimator for this conditional probability is

$$p(e_t|e_{t-n}^{t-1}) = \frac{c(e_t|e_{t-n}^{t-1})}{\sum_{e \in \mathcal{A}} c(e|e_{t-n}^{t-1})} \tag{3}$$

where $c(e_t|e_{t-n}^{t-1})$ denotes the number of times symbol e_t has been observed following context e_{t-n}^{t-1}, computed over the entire observation sequence e_1^{t-1}. To estimate the probability of unobserved events, we incorporate a smoothing approach [14], so that recursively,

$$p(e_t|e_{t-n}^{t-1}) = \begin{cases} \alpha(e_t|e_{t-n}^{t-1}) & \text{for } c(e_t|e_{t-n}^{t-1}) > 0 \\ \gamma(e_{t-n}^{t-1})\, p(e_t|e_{t-n+1}^{t-1}) & \text{otherwise.} \end{cases} \tag{4}$$

In Equation 4, $\alpha(\cdot|\cdot)$ is defined as follows. It is used as long as the sequence e_{t-n}^{t} has previously been observed at least once. Alternatively, the conditional probability is recursively evaluated using a function $\gamma(\cdot)$ and a lower order estimation $p(e_t|e_{t-n+1}^{t-1})$.

As employed in [8], $\alpha(\cdot|\cdot)$ and $\gamma(\cdot)$ are defined as

$$\gamma(e_{t-n}^{t-1}) = \frac{d(e_{t-n}^{t-1})}{\sum_{e \in \mathcal{A}} c(e|e_{t-n}^{t-1}) + d(e_{t-n}^{t-1})} \tag{5}$$

$$\alpha(e_t|e_{t-n}^{t-1}) = \frac{c(e_t|e_{t-n}^{t-1})}{\sum_{e \in \mathcal{A}} c(e|e_{t-n}^{t-1}) + d(e_{t-n}^{t-1})} \tag{6}$$

where $d(e_{t-n}^{t-1})$ denotes the number of distinct symbols observed as continuations of context e_{t-n}^{t-1}. Intuitively, as $d(\cdot)$ increases, more emphasis is placed on shorter contexts when estimating unobserved symbol probabilities.

Since the N-gram model is based on an alphabet of discrete symbols, we quantise our continuous-valued feature vectors prior to learning this model. This is achieved using online k-means clustering, described in Section 3.2.

2.3 Repetition detection

We propose the use of a repetition detection algorithm to inform predictions in conjunction with autoregressive and N-gram approaches. To incorporate information on long-term

structure as described in Section 1, the similarity between feature vector sequences is computed during the prediction process. As incorporated in [20], the approach uses a self-distance matrix (SDM). Given observations $\mathbf{v}_1, \ldots, \mathbf{v}_{t-1}$, the SDM $\mathbf{D}$ is defined as $[\mathbf{D}]_{i,j} = d(\mathbf{v}_i, \mathbf{v}_j)$, with $1 \leq i, j < t$. As proposed in [7], for the distance function $d(\cdot, \cdot)$ we use the cosine distance,

$$d(\mathbf{v}_i, \mathbf{v}_j) = 0.5 \left(1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}\right). \tag{7}$$

Assume a predefined sequence comparison length L. We use the SDM to consider all alignments between past sequences $\mathbf{v}_{s-L}, \ldots, \mathbf{v}_{s-1}$ and the most recently observed feature vectors $\mathbf{v}_{t-L}, \ldots, \mathbf{v}_{t-1}$, with $L < s < t$. Comparing vector-wise with the most recently observed feature vectors, the past sequence with minimal average distance is selected as the conjectured repeated sequence. This sequence is used for prediction, assuming that $\mathbf{v}_t \approx \mathbf{v}_s$. With $L < t \leq T$, the tth prediction $\mathbf{w}_t$ is obtained using index p of past vector $\mathbf{v}_p$, with

$$p = \underset{L < s < t}{\operatorname{argmin}} \left\{ d_\mu(s, t) \right\}, \tag{8}$$

where $d_\mu(s, t)$ denotes the average distance between two subsequences of length L,

$$d_\mu(s, t) = \frac{1}{L} \sum_{\ell=1}^{L} [\mathbf{D}]_{s-\ell, t-\ell}. \tag{9}$$

Computing the entire sequence of predictions has polynomial time complexity against the total sequence length T, since each prediction at step t requires $O(T)$ operations. We observe that using beat-synchronous features results in an average sequence length of approximately 650, for the data set of popular music chosen for evaluation. Therefore scalability is not thought to restrict the algorithm's utility, for music signals with similar duration to those in the data set. Furthermore, it is possible to deal with longer music signals by imposing a maximum size on the SDM, discarding observations which fall outside a specified history limit.

2.4 Combining multiple predictors

To combine the predictions generated by the SDM and N-gram approaches, we propose a linear weighting scheme based on estimated variance of error [1]. For a set of M predictors, define the tth prediction by each predictor $\mathbf{v}_t^i$, with $1 \leq i \leq M$. Define the true value of the tth vector to be $\mathbf{v}_t^*$. We assume an observation model where predictions $\mathbf{v}_t^i$ are the sum of observations $\mathbf{v}_t^*$ and an error term ϵ_t^i,

$$v_{t,u}^i = v_{t,u}^* + \epsilon_{t,u}^i \tag{10}$$

where indices u denote vector components, with $1 \leq u \leq k$. We assume components $\epsilon_{t,u}^i$ to be normally distributed, with variance $\sigma_{i,u}^2$. Using H predictions as samples, the variance of the error $\sigma_{i,u}^2$ can be estimated as

$$\hat{\sigma}_{i,u}^2 = \frac{1}{H-1} \sum_{h=1}^{H} \left(v_{t-h,u}^i - v_{t-h,u}^*\right)^2. \tag{11}$$

Because the error is assumed to be normal, we have $p(v_{t,u}^i | v_{t,u}^*) = \mathcal{N}(v_{t,u}^*, \sigma_{i,u}^2)$. Using Bayes' theorem, we have

$$p(v_{t,u}^* | v_{t,u}^i) = \frac{p(v_{t,u}^i | v_{t,u}^*) \, p(v_{t,u}^*)}{p(v_{t,u}^i)}. \tag{12}$$

If we assume the ratio of $p(v_{t,u}^i)$ and $p(v_{t,u}^*)$ is non-informative, we then have $p(v_{t,u}^i | v_{t,u}^*) = p(v_{t,u}^* | v_{t,u}^i)$. We further assume independence between predictors and denote $\beta_{i,u} = 1/\sigma_{i,u}^2$ for notational convenience. Then, the distribution of $v_{t,u}^*$ can be expressed as

$$\begin{aligned} p(v_{t,u}^* | v_{t,u}^1, \ldots, v_{t,u}^M) &= \prod_{i=1}^{M} \mathcal{N}\left(v_{t,u}^*; v_{t,u}^i, \frac{1}{\beta_{i,u}}\right) \\ &= \mathcal{N}\left(v_{t,u}^*; \frac{\sum_{i=1}^{M} \beta_{i,u}}{\sum_{j=1}^{M} \beta_{j,u}}, \frac{1}{\sum_{i=1}^{M} \beta_{i,u}}\right). \end{aligned} \tag{13}$$

Given all predictions, the expected value of $v_{t,u}^*$, $\mathbf{E}[v_{t,u}^*]$ is then the weighted sum

$$\mathbf{E}\left[v_{t,u}^*\right] = \frac{\sum_{i=1}^{M} \beta_{i,u} \, v_{t,u}^i}{\sum_{j=1}^{M} \beta_{j,u}}. \tag{14}$$

Equation 14 describes the weighting scheme used to combine multiple predictions. Note that values $\beta_{i,u}$ describe the precision of prediction method i, estimated over prediction history of length H.

3. METHOD

The data set used for evaluation consists of 180 mono audio tracks of songs by The Beatles, with each track sampled at 44.1kHz [13].

3.1 Feature extraction

We extract beat-synchronous chroma features using the approach and implementation described in [5]. These chroma features are based on the mapping of FFT bins to twelve pitch class components, using phase derivatives to reduce the influence of non-tonal components present in the spectrum. Chroma frames are based on an FFT window size of 2048 with 75% overlap. This approach compensates for mistuning by computing the optimal alignment between frequency peaks and chroma bins over the entire signal.

Furthermore, we extract beat-synchronous MFCCs, using the approach and implementation described in [18]. The MFCCs are based on an FFT window size of 512 with 50%

[1] The method is similar in spirit to aggregation methods reviewed in [21].

overlap. The filter bank consists of 13 linearly spaced filters and 27 log spaced filters. We extract the 12 first cepstal coefficients, omitting the d.c. coefficient. Beat-synchronous MFCCs are then obtained by computing mean feature values within each beat onset interval, applying the same onset intervals used for chroma feature extraction.

The beat onset times are estimated using the code and approach described in [5]. In terms of the causal prediction problem which this work addresses, we note that the method's application of dynamic programming is non-causal. In this work, we treat the beat tracking routines as an oracle for obtaining beat onset times.

3.2 Online clustering

To obtain discrete symbols for the N-gram predictor, we quantise observed feature vectors using online k-means clustering. As described in [8], an initial codebook of K centroids $\mu_1, \ldots, \mu_K$ is constructed according to the first Q observed symbols. Thereafter, upon observing feature $\mathbf{v}_t^*$, the closest centroid

$$\mu_t = \underset{1 \leq k \leq K}{\operatorname{argmin}} \left\{ \|\mathbf{v}_t^* - \mu_k\|^2 \right\} \tag{15}$$

is updated according to

$$\mu_t := \mu_t + \eta(\mathbf{v}_t^* - \mu_k). \tag{16}$$

In our evaluation, we set $Q = K$. A hold-out set of 60 random songs is formed. A learning factor of $\eta = 0.4$ is determined, based on MFCC and chroma prediction performance and using the described data set with a fixed codebook size of $K = 64$. For fixed $\eta = 0.1$, alternative strategies for codebook construction were evaluated, involving initialisation to held out data. However, these revealed no compelling improvement over the aforementioned method, in terms of N-gram prediction performance.

For the N-gram predictor, prediction proceeds causally, so that after the tth prediction, N-gram probabilities are updated to include the actually observed symbol e_t^* and its context e_{t-n}^{t-1}. The N-gram predictor is learned using only observations from the target song. Given the average length of 650 symbols per song, we estimate the required codebook size to be in the order of $\sqrt{650} \approx 25$ symbols. Considering that the N-gram model incorporates a smoothing scheme (cf. Equation 4), we set the Markov order to constant $n = 5$, observing similar prediction performance for $n = 2$. Using the held-out data set of 60 songs, we set respective SDM prediction lengths $L = 22$, $L = 36$, which maximise prediction performance for chroma and MFCCs.

3.3 Performance statistics

The statistics used for evaluation are the sum of squares error (SSE), the Jensen-Shannon divergence (JSD) and the absolute deviation (AD). The SSE for the tth prediction is computed as

$$\mathrm{SSE}(\mathbf{v}_t, \mathbf{v}_t^*) = \|\mathbf{v}_t - \mathbf{v}_t^*\|^2. \tag{17}$$

The JSD is a symmetrised version of the Kullback-Leibler divergence. It is computed as

$$\mathrm{JSD}(\mathbf{v}_t \| \mathbf{v}_t^*) = \frac{1}{2} KL\left(\mathbf{v}_t, F\right) + \frac{1}{2} KL\left(\mathbf{v}_t^*, F\right) \tag{18}$$

where $KL(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence and F is defined as

$$F = \frac{1}{2}\left(\mathbf{v}_t + \mathbf{v}_t^*\right). \tag{19}$$

Finally, the absolute deviation is computed as

$$\mathrm{AD}(\mathbf{v}_t, \mathbf{v}_t^*) = \|\mathbf{v}_t - \mathbf{v}_t^*\|_1 \tag{20}$$

where $\|\cdot\|_1$ denotes the ℓ^1-norm.

We compute the statistics for all predictions and average over predictions in the entire data set. For example, the average sum of squares error $\mathrm{SSE}_\mu$is computed as

$$\mathrm{SSE}_\mu = \frac{1}{T} \sum_{t=1}^{T} \mathrm{SSE}(\mathbf{v}_t, \mathbf{v}_t^*). \tag{21}$$

Average prediction results therefore describe vector-wise prediction error and do not account for variability in song duration. We compute 99% confidence intervals on average performance data. Relative to LPC prediction performance, confidence intervals do not exceed 3.4%, 1.8%, 0.2%, in terms of average SSE, JSD and AD, respectively.

4. RESULTS

We evaluate autoregressive, N-gram and SDM predictors. Designating the LPC predictor as a baseline, Figure1 illustrates prediction performance relative to the LPC baseline, in terms of average SSE, JSD, AD. Performance values are expressed as the quotient S/B, where S is the average prediction error of the sample and B is the average prediction error of the LPC baseline.

4.1 Single predictor performance

We first consider the accuracy of individual predictors, with no method of combining them applied. On the left hand side of Figure 1 (a), (b), we include results for four prediction techniques. Based on the assumption of local stationarity, the predictor termed 'Copy' estimates the tth prediction as $\mathbf{v}_t^c = \mathbf{v}_{t-1}^*$. The predictor termed 'LPC' applies the linear predictor described in Equation 2. The predictor termed 'MAR' performs multivariate autoregression according to Equation 1. The predictor termed 'SDM' corresponds to repetition detection using a self-distance matrix, as described in Section 2.3. For both LPC and MAR predictors, all observations $\mathbf{v}_1^*, \ldots, \mathbf{v}_{t-1}^*$ are incorporated into

a least-squares regression [12]. Results are reported for second order LPC models (chroma), third order LPC models (MFCC) and first order MAR models (chroma and MFCC), with orders selected to maximise held-out data performance.

Considering chroma feature prediction in Figure 1 (a), we observe that Copy prediction is significantly outperformed by all remaining predictors, for all evaluated statistics. Observing that the MAR model is outperformed by LPC based prediction, it appears that for the given sequence lengths and the chosen features, it is preferable to assume independence between feature components.

For the considered codebook sizes, the N-gram model is almost consistently outperformed by the LPC predictor. To reduce the error that is due to quantisation alone, we weight predicted feature vectors using the linear combination $(1-\gamma)\,\mathbf{v}^n + \gamma\,\mathbf{v}^c$, where $\mathbf{v}^n$ is the discrete N-gram prediction. Parameter γ is varied within the unit interval, in steps of 0.1. Based on 10×2 cross-validation on the remaining 120 songs, results are reported for $\gamma = 0.4$, which maximises SSE performance for both chroma and MFCC features. In Figure 1, this predictor is termed 'Weighted'.

Considering MFCC feature prediction in Figure 1 (b), we observe that SDM prediction offers less advantage over Copy prediction, compared to chroma prediction.

Turning to the effect of increasing codebook size, we observe that SSE performance improves for chroma predictions. Surprisingly, for MFCC prediction increasing the codebook size adversely affects SSE performance. In both cases, JSD and AD performance degrades when increasing codebook size.

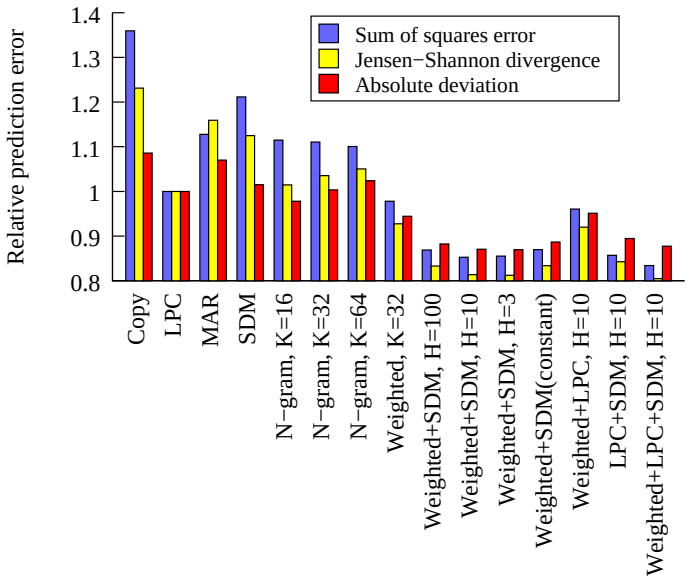
4.2 Combined predictor performance

Results for combinations of predictors are shown on the right hand side of Figure 1 (a), (b). To restrict the parameter space, the evaluation is based on the aforementioned baseline results. Thus, the linear chroma weighting factor is set to $\gamma = 0.4$. Based on Equation 14, a running estimate of precision values β_i is formed using $\min\{H, t-1\}$ preceding predictions.

Results for chroma and MFCC feature prediction reveal that combining SDM with weighted N-gram approaches results in substantial improvement over single predictor performance. The result is largely consistent across the evaluated SSE, JSD and AD statistics. We observe a similar result when combining LPC and SDM predictors. Compared to the latter result, combining LPC, SDM and weighted N-gram predictors further improves performance.

For comparison, a linear and constant weighting scheme was evaluated. As reported in Figure 1 ('constant'), no improvement over history based weighting is obtained using this approach.

(a) Chroma feature prediction



(b) MFCC feature prediction

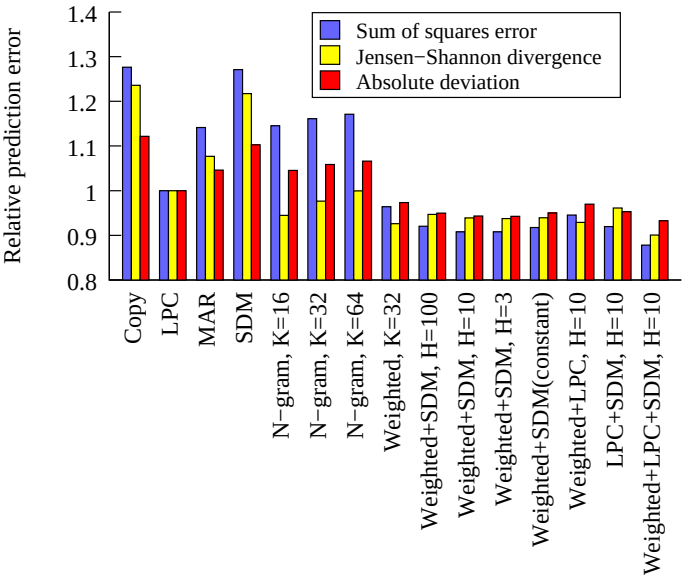


Figure 1. Performance results for chroma and MFCC feature prediction. Parameter K denotes codebook size. Parameter H denotes amount of prediction history used to inform predictor combination. See main text for a description of predictor labels. Absolute chroma performance values for the LPC baseline are 0.0568 (SSE) 0.0882 (JSD) 0.453 (AD). Absolute MFCC performance values for the LPC baseline are 0.893 (SSE) 0.406 (JSD) 2.282 (AD).

Approach	Chroma	MFCC	Average
N-gram (weighted)	5%	5%	5%
LPC + SDM	14%	6%	10%
N-gram (weighted) + SDM	15%	7%	11%
N-gram (weighted) + SDM + LPC	16%	10%	13%

Table 1. Summary of average chroma and MFCC prediction performance. Scores are gains relative to the LPC baseline.

4.3 Summary of results

Table 1 summarises the obtained results. For each statistic, we describe performance gains relative to the LPC baseline, averaged across SSE, JSD and AD statistics.

We observe that using the weighted N-gram approach yields minor improvement over the baseline LPC method. This result is consistent for both chroma and MFCC prediction tasks. A further result concerns the inclusion of the SDM approach: In combination with either weighted N-gram or LPC approaches, we observe average performance gains in excess of 6%. Average chroma prediction performance improves by at least 14%. Furthermore, combining N-gram and SDM predictors yields minor improvement over the analogous LPC and SDM combination.

5. CONCLUSIONS AND FURTHER WORK

In this work, we have considered the problem of causal music prediction using MFCC and chroma features. We have comparatively evaluated the performance of predictors for series of continuous-valued and quantised feature vectors. We have considered how musical parallelism might be harnessed for causal prediction of spectro-temporal features. The prediction approach proposed in this work is based on repetition detection using a self-distance matrix.

For the evaluated statistics, combining the SDM predictor with LPC or N-gram approaches allows substantial improvements in prediction accuracy to be made, compared to the baseline. This suggests that incorporating information on long-term musical structure might have utility for the causal prediction of spectro-temporal features.

Considering the obtained results, we plan investigations to determine the effectiveness of online quantisation, the prerequisite for applying discrete-event models such as the N-gram model used in this work. Furthermore, we aim to perform an evaluation of hierarchical language models based on the N-gram model used in this work. Finally, we aim to consider music prediction from a perceptual perspective, to identify correlates between perceived musical similarity and prediction accuracy.

6. ACKNOWLEDGEMENTS

This work benefited from advice from Andrew Robertson, Adam Stark and Roger Dean. In addition, we would like to thank the anonymous reviewers for their comments.

7. REFERENCES

[1] J. Bello: "Grouping Recorded Music by Structural Similarity," *Proc. ISMIR*, pp. 531–536, 2009.

[2] I. Bent and W. Drabkin: *Analysis. New Grove Handbooks in Music*, Macmillan, London, 1987.

[3] A. Cont: *Modeling musical anticipation: From the Time of Music to the Music of Time*, Ph.D. Thesis, University of California, San Diego, San Diego, 2010.

[4] S. Dubnov: "Unified View of Prediction and Repetition Structure in Audio Signals with Application to Interest Point Detection," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 2, pp.327–337, 2008.

[5] D. Ellis and G. Poliner: "Identifying 'Cover Songs' with Beat-synchronous Chroma Features," *Proc. Intern. Conference on Acoustics, Speech and Signal Processing*, pp. 1429-1432, 2007.

[6] S. Farner, A. Solvang, A. Saebo and U. Svensson: "Ensemble Hand-clapping Experiments Under the Influence of Delay and Various Acoustic Environments," *Journal of the Audio Engineering Society*, Vol. 57, No. 12, pp. 1028–1041, 2009.

[7] J. Foote: "Visualizing Music and Audio Using Self-similarity," *Proc. ACM Intern. Conference on Multimedia*, pp. 77–80, 1999.

[8] A. Hazan: *Musical Expectation Modelling from Audio: A Causal Mid-level Approach to Predictive Representation and Learning of Spectro-temporal Events*, Ph.D. Thesis, Universitat Pompeu Fabra, Barcelona, 2010.

[9] D. Huron: *Sweet Anticipation: Music and the Psychology of Expectation*, The MIT Press, Cambridge, MA, 2006.

[10] B. Jung, J. Hwang, S. Lee, G. Kim, and H. Kim: "Incorporating Co-presence in Distributed Virtual Music Environment," *Proc. ACM Symposium on Virtual Reality Software and Technology*, pp. 206–211, 2000.

[11] F. Lerdahl and R. Jackendoff: *A Generative Theory of Tonal Music*, The MIT Press, Cambridge, MA. 1996.

[12] H. Lütkepohl: *New Introduction to Multiple Time Series Analysis*, Springer, Berlin, 2005.

[13] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler: "OMRAS2 Metadata Project 2009," *Proc. ISMIR*, 2009.

[14] A. Moffat: "Implementing the PPM Data Compression Scheme," *IEEE Transactions on Communications*, Vol. 38, No. 11, pp. 1917–1921, 1990.

[15] J. Paulus and A. Klapuri: "Acoustic Features for Music Piece Structure Analysis," *Proc. Intern. Conference on Digital Audio Effects*, pp. 309–312, 2008.

[16] M. Pearce and G. Wiggins: "Improved Methods for Statistical Modelling of Monophonic Music," *Journal of New Music Research*, Vol. 33, No. 4, pp. 367–385, 2004.

[17] T. Schneider and A. Neumaier: "Algorithm 808: ARfit–A Matlab Package for the Estimation of Parameters and Eigenmodes of Multivariate Autoregressive Models," *ACM Transactions on Mathematical Software*, Vol. 27, No. 1, pp. 58–65, 2001.

[18] M. Slaney: "Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work," *Interval Research Corporation*, 1998.

[19] R. Solomonoff: "A Formal Theory of Inductive Inference: Part 1 and 2," *Inform. Control*, Vol. 7, pp. 224–254, 1964.

[20] A. Stark and M. Plumbley: "Performance Following: Real-Time Prediction of Musical Sequences Without a Score," *To appear in IEEE Transactions on Audio, Speech, and Language Processing*.

[21] V. Vovk: "Competitive On-line Statistics," *Intern. Statistical Review*, Vol. 69, pp. 213–248, 2001.