A MODEL SELECTION TEST FOR FACTORS AFFECTING THE CHOICE OF EXPRESSIVE TIMING CLUSTERS FOR A PHRASE

Shengchen Li, Simon Dixon
Queen Mary University of London
shengchen.li@hotmail.com
s.e.dixon@qmul.ac.uk

Dawn A. A. Black
Radioscape
dawn.black@
radioscape.co.uk

Mark D. Plumbley
University of Surrey
m.plumbley@surrey.ac.uk

ABSTRACT

We model expressive timing for a phrase in performed classical music as being dependent on two factors: the expressive timing in the previous phrase and the position of the phrase within the piece. We present a model selection test for evaluating candidate models that assert different dependencies for deciding the Cluster of Expressive Timing (CET) for a phrase. We use cross entropy and Kullback Leibler (KL) divergence to evaluate the resulting models: with these criteria we find that both the expressive timing in the previous phrase and the position of the phrase in the music score affect expressive timing in a phrase. The results show that the expressive timing in the previous phrase has a greater effect on timing choices than the position of the phrase, as the phrase position only impacts the choice of expressive timing in combination with the choice of expressive timing in the previous phrase.

1. INTRODUCTION

In classical music, performers vary the lengths of beats throughout a performance while keeping the overall beat rate. Such small variations of beat timing are known as expressive timing. Expressive timing contributes to the formation of expressiveness in classical music. Research into expressive timing shows that the expressive timing within a phrase is not randomly distributed but similar timing profiles are used across different phrases. There are various investigations about how such similar timing profiles can be found and how such common timing profiles are used by performers.

It is common in the literature to cluster the expressive timing in performed classical piano music into different types, with various temporal units used. For example, Repp [1] uses principal component analysis to analyse the commonalities and differences in performances of a Chopin Étude. Spiro et al. [2] use a self-organising map to cluster the expressive timing within a bar and investigate how the clusters of expressive timing are distributed. With model selection tests, Li et al. [3] demonstrate that clustering the expressive timing within a phrase is helpful for analysing expressive timing. Moreover, Li et al. [3] also introduce a method to cluster the expressive timing within a phrase by using a Gaussian mixture model. In this paper, we make use of model selection tests to show how the choice of Cluster of Expressive Timing (CET) is possibly affected.

There have been a few attempts to determine how expressive timing varies in a segment of performance. In [4] and [5], Widmer et al. discuss how expression in performed music is formed when the musical score is given. Their basic idea for expressiveness synthesis is to render each phrase using expressive gestures extracted from performances of similar phrases in a training database. In [4], the authors suggest that a dynamic Bayesian network may be used for expressiveness synthesis, in which case, the expressive timing in the previous parts of performance may affect the expressive timing in later parts. Similarly, in [6], Todd points out that parabolic curves can be used for fitting tempo variations across different levels of music structure. This suggests that tempo variations within a phrase can be affected by expressive timing in previous parts. Moreover, in the rule based system from KTH [7], the expressive timing is affected by both the music score and the sequence of expressive timing. As a summary of the works mentioned above, the music score and expressive timing in previous parts may affect the current choice of expressive timing.

In this paper, we examine two possible factors affecting the choice of CET for a phrase: the position of the phrase and the CET used in the previous phrase. In particular, we use model selection tests to demonstrate how the CET in a phrase is affected by both the CET used in the previous phrase and the position of the phrase in the musical score. We propose four Bayesian graphical models that assert different relationships between the CET used in a particular phrase, the CET used in the previous phrase and the position of the phrase. Then we design a model selection test to evaluate how well the candidate models predict the use of CETs. As the candidate models have different structures, we use cross entropy and Kullback Leibler (KL) divergence to evaluate the resulting models. Cross entropy and KL divergence are both derived from information theory and can evaluate models in different model spaces.

To obtain the CET distribution in this analysis, we follow the procedure developed in previous work [3], and use the same database: two Chopin Mazurkas (Op.24/2 and Op.30/2) and *Islamey* by Mily Balakirev [8]. In each candidate piece, the phrase lengths are identical throughout the piece. In addition, the beat timing for the two Mazurkas is provided in the Mazurka database, which is used in various works [2, 9] by the CHARM group. The number of CETs varies from piece to piece according to our published methodology [3].

This paper is organised in the following way: we firstly introduce how the expressive timing patterns within a phrase are clustered. Then we observe how the CETs are distributed across different performers throughout a piece of music. Next we introduce the candidate models in this paper. Then we present the evaluation of the candidate models, followed by a discussion and conclusion.

2. CLUSTERING OF EXPRESSIVE TIMING

In this section, we describe how expressive timing is clustered in this work. For two Chopin Mazurkas, the tempo data is provided by the database. For *Islamey*, only beat timing is provided. We now introduce how we convert beat timing to tempo data for *Islamey* database. If we use t_i to represent the beat timing of the ith beat, for a piece of music that has n beats, the beat timing can be represented as $t_1, t_2, \ldots, t_n, t_{n+1}$ where t_{n+1} represents the ending time of the last beat in the piece. We use the reciprocal of the inter beat interval to represent tempo at the beat level (i.e. $\tau_i = \frac{1}{t_{i+1}-t_i}$). As mentioned above, the candidate pieces have constant phrase lengths throughout the piece, thus the expressive timing within the ith phrase that has w beats can be represented as $\mathbf{T_i} = (\tau_1, \tau_2, \ldots, \tau_w)$. By the expectation maximisation method, we can fit the distribution of expressive timing within a phrase to a Gaussian mixture model such that:

$$p(\mathbf{T_i}) = \sum_{a=1}^{A} \pi_a \mathcal{N}(\tau_i|\mu_a, \mathbf{\Sigma}_a^{full}), \tag{1}$$

where there are A clusters available, each with mean μ_a, covariance $\mathbf{\Sigma}_a^{full}$, and weight π_a for index a. If we use T_i^* to represent the CET that the expressive timing in phrase i belongs to, we have:

$$T_i^* = arg_a \max \pi_a \mathcal{N}(\tau_i|\mu_a, \mathbf{\Sigma}_a^{full}). \tag{2}$$

As discussed in previous work [3], the optimum number of CETs for a phrase differs from piece to piece. Using cross validation tests, the optimum number of CETs for the candidate pieces was found to be 2 clusters for *Islamey*, 8 clusters for Mazurka Op.24/2 and 4 clusters for Mazurka Op.30/2 [3].

Suppose that there are n phrases in a candidate piece of music and there are m performances in the database. If we use a vector to represent the clusters of expressive timing used for each phrase in performance j, we have $\mathbf{P_j^*} = (T_{1j}^*, T_{2j}^*, \ldots, T_{nj}^*)$. Thus we can use a matrix $\mathbf{P}^*$ whose row is $\mathbf{P_j^*}$ for performer j to represent the clusters of expressive timing used in each phrase for all performers. For easier observation, we convert matrix $\mathbf{P}^*$ to a diagram so that each element in $\mathbf{P}^*$ is represented by a colour block according to the cluster of expressive timing used. This type of diagram is called a Tempo Variegation Map (TVM) [10]. In Figure 1, we give a TVM for Mazurka Op.24/2 as an example. In this diagram, each row represents a performance of the Mazurka and each column represents a phrase. Each colour block represents a CET used in a phrase. The colours of blocks are selected according to the centroids of the CETs, with similar colours representing those clusters whose centroids are similar. By observing the distribution of the CETs, we propose the candidate models in this work.

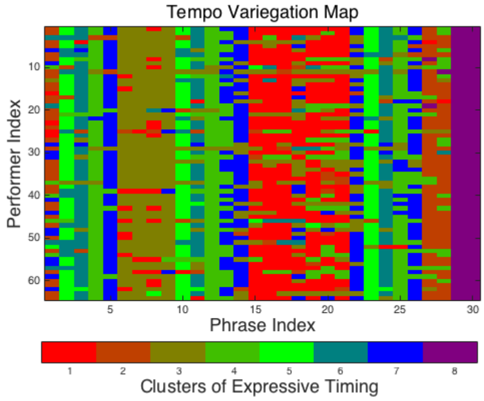


Figure 1: An example of a Tempo Variegation Map (TVM) for Mazurka Op.24/2.

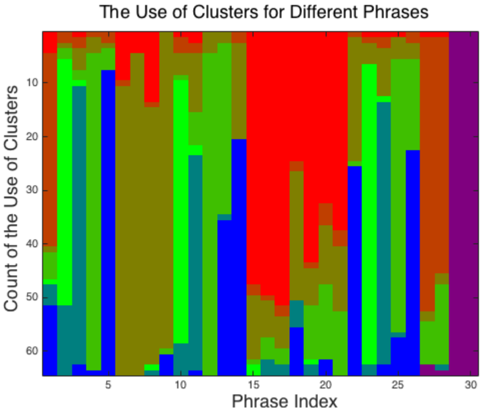


Figure 2: The distribution of CET used by all the performers for each phrase in Mazurka Op.24/2. The colours match those in Figure 1.

3. CANDIDATE MODELS

In this section, we introduce the four candidate models we propose according to our observations of the TVMs. To illustrate the observations for the candidate models, we use Figure 1 as an example. Then we introduce some regularities of the distribution of CETs and give the mathematical descriptions of the candidate models.

In Figure 1, we can see that for some phrases, the use of CET agrees across different performers. If we count the frequency of each CET for each phrase in a performance, we obtain Figure 2. In this figure, we see how the frequency of CETs differs from phrase to phrase, thus we propose the *positional model* (PM), which asserts that the position of the phrase in the music score affects the choice

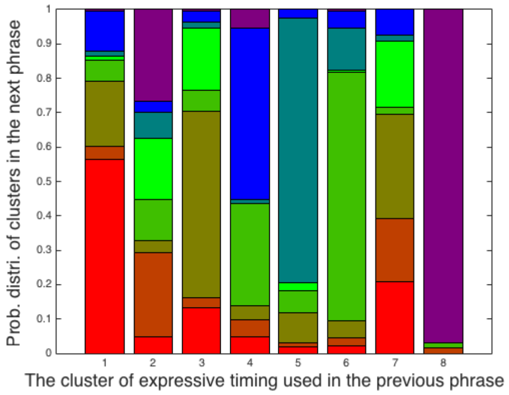


Figure 3: The relative frequency of CETs used by all the performers after a specific cluster is used in Mazurka Op.24/2. The colours match those in Figure 1.

of CET for a phrase.

Furthermore, if we observe Figure 1 again, we can see that some CETs are likely to be followed by a particular CET. For example, in Figure 1, cluster 5 is likely to be followed by cluster 6. If we visualise the relative frequency of each CET that appeared after another CET, we obtain Figure 3 that visualises the different distribution of CETs after a particular CET is used in the previous phrase. As a result, we propose the *sequential model* (SM), which asserts that the choice of CET for a phrase is affected by the CET used in the previous phrase.

Beside the positional model and the sequential model, we propose two other candidate models: a joint model and an independent model. The *joint model* (JM) asserts that the choice of CET for a phrase is affected by both the position of phrase and the CET used in the previous phrase. The *independent model* (IM) is a reference model which asserts that neither the position of phrase nor the CET in the previous phrase have any effect on the choice of CET for the next phrase.

In the candidate models, there are three variable parameters: the CET used in a particular phrase (T_i^*), the position of the phrase (β) and the CET used in the previous phrase (T_{i-1}^*). All candidate models are Bayesian graphical models that can be extended to a joint probability distribution of the parameters in the candidate models (namely $p(T_{i-1}^*, T_i^*, \beta)$). In the model selection test we use a cross-validation method to randomly select rows in $\mathbf{P}^*$ to form a testing dataset, with the remaining data in $\mathbf{P}^*$ forming the training dataset. We use five-fold cross validation to evaluate the candidate models. To remove the possible effects of the random train/test split, we repeat the five-fold cross validation tests several times.

Each training dataset is trained for finding $p(T_{i-1}^*, T_i^*, \beta)$ in the testing dataset. Then we evaluate how successfully $p(T_{i-1}^*, T_i^*, \beta)$ from the testing dataset is predicted according to the training dataset. The results derived from different formations of testing and training datasets are averaged. In some cases, certain combinations of $(T_{i-1}^*, T_i^*, \beta)$ may be absent in the training datasets but appear in the testing datasets. This will cause a problem of zero probability [11, Ch.17]. We use Bayesian estimation to learn the parameters in the candidate models to prevent the zero probability problem, which adds a small count to all probabilities [11, Ch.17]. For example, if there are x_1 samples such that $X = 1$ in a database that has x samples, the probability of $X = 1$ is defined by Bayesian estimation as:

$$p(X = 1) = \frac{x_1 + \frac{1}{x}}{x + 1}. \tag{3}$$

With the rule of multiplication for probability (if event A and event B are independent, $p(A, B) = p(A)p(B)$), Equations (4), (5), (6), and (7) define how $p(T_{i-1}^*, T_i^*, \beta)$ is calculated according to the Independent Model (IM), Positional Model (PM), Sequential Model (SM) and Joint Model (JM) respectively.

$$p_{\mathrm{IM}}(T_{i-1}^*, T_i^*, \beta) = p(T_{i-1}^*) \times p(T_i^*) \times p(\beta) \tag{4}$$

$$p_{\mathrm{PM}}(T_{i-1}^*, T_i^*, \beta) = p(T_{i-1}^*) \times p(T_i^* | \beta) \tag{5}$$

$$p_{\mathrm{SM}}(T_{i-1}^*, T_i^*, \beta) = p(T_{i-1}^* | T_i^*) \times p(\beta) \tag{6}$$

$$p_{\mathrm{JM}}(T_{i-1}^*, T_i^*, \beta) = \frac{\mathit{Count}(T_{i-1}^*, T_i^*, \beta) + \frac{1}{N}}{N + 1} \tag{7}$$

where N is the number of samples.

4. MODEL EVALUATION

The parameters in the candidate models are trained with the training datasets. Then we design a set of model selection tests to evaluate the candidate models to investigate how CETs are affected. In previous work [3] we demonstrated that model selection tests can be used for expressive timing analysis, using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). However, the AIC and BIC are designed to compare candidate models in the same model space [12, Ch.2-3], so in this paper, we use a more general method that compares the joint probability of $(T_{i-1}^*, T_i^*, \beta)$ in both training and testing datasets. The parameters used for model evaluation include cross entropy and KL divergence.

To distinguish the joint probability distribution in the training datasets and the testing datasets, we use $q(T_{i-1}^*, T_i^*, \beta)$ to represent the joint probability in the training dataset and $p(T_{i-1}^*, T_i^*, \beta)$ to represent the joint probability distribution in the testing dataset. For simplicity, we have $q_i \equiv q_{klm} \equiv q(T_{i-1}^* = k, T_i^* = l, \beta = m)$ and $p_i \equiv p_{klm} \equiv p(T_{i-1}^* = k, T_i^* = l, \beta = m)$.

The cross entropy between P and Q measures how many bits on average are required to code a symbol in P if we have a coding system whose probability distribution of symbols is given by Q. The cross entropy [13] for a distribution of n symbols is defined as:

$$H_{\text{Cross}}(P, Q) = -\sum_{i=1}^{n} p_i \log_2(q_i). \tag{8}$$

KL divergence, or relative entropy, is an indicator of how different a probability distribution Q is when compared to probability distribution P. The KL divergence is not a strict distance measurement, due to the fact that it is not symmetric ($KL_{Div}(P,Q) \not\equiv KL_{Div}(Q,P)$). The KL divergence is equivalent to the difference between cross entropy and the entropy of the testing dataset, as Equation (9) shows. In other words, KL divergence measures how efficient the coding system optimised for Q is for coding P.

$$\begin{aligned} KL_{Div}(P,Q) &= \sum_{i=1}^{n} p_i \log_2(\frac{q_i}{p_i}) \\ &= \sum_{i=1}^{n} \{p_i \log_2(q_i) - p_i \log_2(p_i)\} \\ &= H_{\text{Cross}}(P,Q) - H(P) \end{aligned} \tag{9}$$

5. RESULTS

In this section, we use two model selection criteria: cross entropy (defined in Equation (8)) and KL divergence (defined in Equation (9)). The model selection tests are applied with three candidate pieces: *Islamey* and two Chopin Mazurkas (Op.24/2 and Op.30/2). Following the method of Li et al. [3], the numbers of CETs used for analysis are 2, 8 and 4 for *Islamey*, Chopin Mazurka Op.24/2 and Op.30/2 respectively.

Model Criterion	IM	PM	SM	JM
Cross Entropy	7.25	7.71	7.12	**6.88**
KL Divergence	1.00	1.46	0.88	**0.63**

(a) *Islamey*

Model Criterion	IM	PM	SM	JM
Cross Entropy	10.63	13.31	9.69	**7.74**
KL Divergence	4.22	6.90	3.24	**1.36**

(b) Chopin Mazurka, Op.24/2

Model Criterion	IM	PM	SM	JM
Cross Entropy	5.80	6.60	5.60	**4.92**
KL Divergence	1.69	2.49	1.49	**0.81**

(c) Chopin Mazurka, Op.30/2

Table 1: Model evaluation of the candidate models that assert different dependencies on the CET used in a phrase. Both model selection criteria use a smaller value to indicate better model performance. The IM, PM, SM and JM are defined in Section 3. The bold value indicates the best performance of the candidate models.

For all candidate pieces, we use five-fold cross-validation to test the candidate models. For a single experiment, we select the data from 20% of performances in our database randomly to form the testing dataset and the remaining 80% of performances forms the training dataset. The experiment is repeated 100 times to mitigate the possible effects of randomness in forming testing and training datasets. The results of the model selection criteria for evaluating how well the testing datasets are predicted are then averaged to obtain the final results.

Model Criterion	IM	PM	SM	JM
Cross Entropy	7.47	9.07	**7.27**	9.05
KL Divergence	0.92	2.56	**0.77**	2.52

(a) *Islamey*

Model Criterion	IM	PM	SM	JM
Cross Entropy	11.07	13.72	**11.01**	11.06
KL Divergence	4.26	6.90	**4.15**	4.39

(b) Chopin Mazurka, Op.24/2

Model Criterion	IM	PM	SM	JM
Cross Entropy	**6.11**	6.60	6.29	6.56
KL Divergence	**1.80**	2.22	1.96	2.13

(c) Chopin Mazurka, Op.30/2

Table 2: Average model selection criteria for a training dataset of only one performance. The IM, PM, SM and JM are defined in Section 3. The bold value indicates the best performance of the candidate models.

In Table 1, we present how well the candidate models predict the testing dataset on average. According to the cross entropy and the KL divergence, the joint model is the best model among the candidate models. The sequential model is the second best model. The positional model is even worse than the independent model.

Data-size robustness means how much the model performance drops when a very limited amount of training data is available. The data-size robustness is a property of the candidate model. In this paper we compare the results obtained using 80% of performances for training (Table 1) with those obtained using only one performance for training (Table 2).

From Table 2, we notice that the data-size robustness of candidate models varies for different candidate pieces. For *Islamey* and Mazurka Op.24/2, the best model in terms of data-size robustness is the sequential model. For Mazurka Op.30/2, which has a training dataset of only 7 phrases, compared with 39 and 29 phrases for *Islamey* and Mazurka Op.24/2 respectively, no model performs better than the baseline independent model. Between the sequential model and the joint model, which have the lowest cross entropy and KL divergence in the model selection tests, the sequential model is more data-size robust.

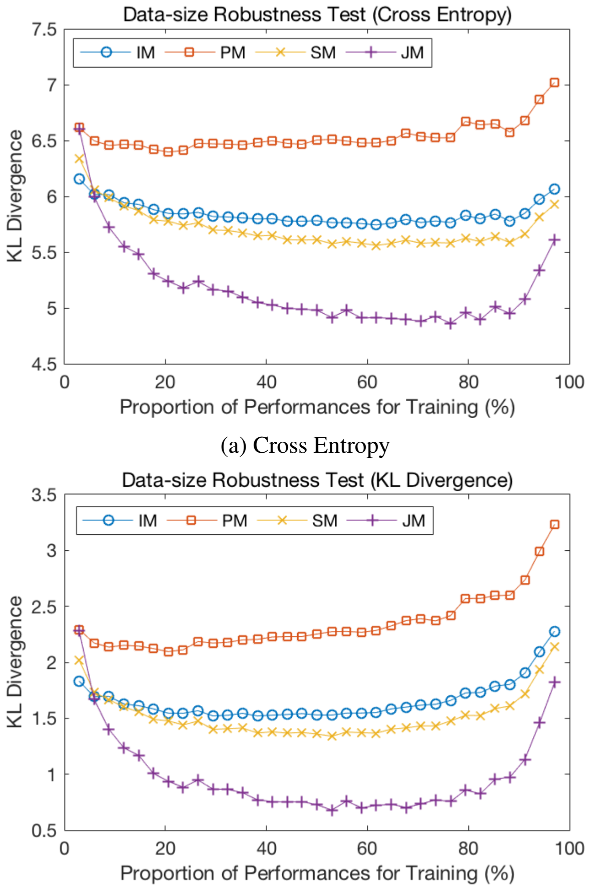
In summary, the joint model is the best model when a reasonable amount of data is available for training. When the amount of training data is limited, the sequential model is preferred as it is more data-size robust with acceptable results. Based on the model selection tests, we demonstrate that both the position of phrase in music score and the CET used in the previous phrase affect the decision of CET for a phrase. The CET in the previous phrase has a greater effect than the position of the phrase, which only affects the choice of CET for a phrase jointly with the CET used in the previous phrase.

6. DISCUSSION

6.1 Comparison of model selection criteria

The results for training with 80% of performances (Table 1) are in some cases worse than those obtained by training on only one performance (Table 2). This fact suggests that using only 1 performance for training results in a better model than using 80% of performances for training in these cases, which conflicts with the intuition that having more data for training usually results in a better model. As a result, we investigate the data-size robustness in more detail.

In Figure 4, we show how cross entropy and KL divergence of candidate models vary with the proportion of performances used for training. Because of the page limitation, we show results only for Chopin Mazurka Op.30/2 as an example in Figure 4. The other pieces give similar results.



(a) Cross Entropy

(b) KL Divergence

Figure 4: Model evaluation as a function of the percentage of performances used for training. The test piece is Chopin Mazurka Op.30/2. A larger number indicates a poorer model.

Observing Figure 4b, the KL divergence start to increase when more than about 70% of performances are used for training. This leads to a higher KL divergence for training sets of 80% of performances than for only one performance for training (compare Tables 1 and 2). In addition, the distribution of testing data can be heavily biased such that even a good model may have a high KL divergence. As a result, model selection tests with KL divergence may be affected by the bias of the testing dataset, whereas the cross entropy test is less affected by bias in the testing dataset.

Usually, if there are more data available for training, we expect the resulting model to be better. However, in Figure 4, the curves of KL divergence and cross entropy are not as monotonic as we expect. If we calculate the first order difference of the cross entropy and KL divergence changes, we find KL divergence has a lower zero-crossing rate than the cross entropy ($p = 0.0161$). In other words, the curves of KL divergence are "smoother" (mathematically speaking, more monotonic) than the curves of cross entropy. The non-monotonic changes in both KL divergence curves and cross entropy curves may be caused by the randomness of the dataset as we have only tried 100 out of the possible 10^{15} formations of the training dataset in this experiment.

Comparing the two model selection criteria used, KL divergence appears to be less sensitive to the randomness of testing data but more sensitive to bias in the testing data, whereas cross entropy is sensitive to the randomness of testing data but is less sensitive to bias. As both model selection criteria agree on the ranking of results, the conclusions in this paper should be robust against the effects of both the randomness and bias of the testing datasets.

6.2 Model complexity

Despite our results showing the rank of candidate model as joint model, sequential model, independent model and positional model, the model complexity of the candidate models do not follow the same order. As there are 2, 8 and 4 CETs chosen in the analysis for *Islamey*, Chopin Mazurka Op.24/2 and Op.30/2 respectively, the number of parameters to be trained in candidate models are listed in Table 3. The number of parameters are decided by the number of phrases in the pieces as well, which is also listed in Table 3. We find that despite the high complexity, the joint model outperforms the other models. The sequential model is more data-size robust due to lower complexity. The results of the model selection tests presented in this paper support the claim that the choice of CET in a phrase is primarily affected by the CET used in the previous phrase. Moreover, the position of phrase in the score only affects the CET in a phrase jointly with the CET used in the previous phrase. Further investigations are required to understand why the positional model alone performs worse than the baseline independent model.

6.3 Future work

In this paper, we present a model selection test that investigates how the CET used in a phrase is affected by the CET used in the previous phrase and the position of a phrase. However, with the same methodology, we can demonstrate how other musical features affect the choice of CETs. The

Pieces	# of phrase	IM	PM	SM	JM
Islamey	40	2	80	4	160
Op.24/2	30	8	240	64	1920
Op.30/2	8	4	32	16	128

Table 3: Number of parameters learnt in candidate models. Abbreviations of models are defined in Section 3.

position of phrase is a simplistic concept which gives little insight into the reasons for expressive choices. There are multiple features in a phrase related to the melody, harmony and rhythm which are likely to influence the performer's choices. For both expressiveness synthesis and musicology research, it would be interesting to investigate further which factors derived from the musical score affect the choice of CET.

Likewise, the exclusive use of the previous CET is a simplification of possible longer term temporal dependencies which may exist between expressive choices, including even non-causal (i.e. planned) relationships with future choices. While the local context is likely to have the largest influence on immediate choices, it would be naive to assume that expert musician's timing choices can be modelled by a simple first-order process.

This research is based on a clustering method proposed in previous work [3]. This method has strong restrictions: the phrase length throughout the candidate piece must be constant, and the number of CETs used varies according to the candidate piece. These restrictions prevent the immediate application of the proposed method to a wider range of performances and a larger dataset using the current clustering algorithm. Further research is required to extend the methods to more general scenarios. Finally, applying the experiments in this paper with generalised model selection methods, such as AIC and BIC, may demonstrate how the model complexity affects the model selection process.

7. CONCLUSIONS

In this paper, we presented a model selection test that investigates how the Cluster of Expressive Timing (CET) is chosen according to the position of the phrase and the CET used in the previous phrase.

We proposed four candidate models that assert different dependencies of the CET used for a particular phrase. We evaluated the four candidate models with KL divergence and cross entropy. The results of the model selection showed that the joint model is the most reasonable model for selecting the cluster of expressive timing for a phrase. However, if there are only very limited data available, the sequential model should be used, owing to its lower complexity.

Hence we have shown that both the CET used in the previous phrase and the position of the phrase affect the selection of CET for a phrase. The sequence of clusters has a greater effect than the position of phrases for selecting the cluster of expressive timing for a phrase. The position of the phrase, on the other hand, only affects the choice in combination with the CET sequences.

8. REFERENCES

[1] B. H. Repp, "A microcosm of musical expression. I. Quantitave analysis of pianists' timing in the initial measures of Chopin's Etude in E major," *The Journal of Acoustical Society of America*, vol. 104, pp. 1085 – 1100, 1998.

[2] N. Spiro, N. Gold, and J. Rink, "The form of performance: Analyzing pattern distribution in select recordings of Chopin's Mazurka op. 24 no. 2," *Musicae Scientiae*, vol. 14, no. 2, pp. 23–55, 2010.

[3] S. Li, D. A. A. Black, and M. D. Plumbley, "Model analysis for intra-phrase tempo variations in classical piano performances," in *Proceedings of Computer Music Multidisciplinary Research (CMMR'15)*, 2015.

[4] G. Widmer, S. Flossmann, and M. Grachten, "YQX plays Chopin," *AI Magazine*, vol. 31, no. 3, pp. 23–34, 2010.

[5] A. Tobudic and G. Widmer, "Relational IBL in music with a new structural similarity measure," in *Proceedings of the 13th International Conference on Inductive Logic Programming (ILP'03)*. Springer, 2003, pp. 365–382.

[6] N. P. M. Todd, "The dynamics of dynamics: A model of musical expression," *Journal of Acoustical Society of America*, vol. 91, pp. 3540–3550, 1992.

[7] A. Friberg, R. Bresin, and J. Sundberg, "Overview of the KTH rule system for musical performance," *Advances in Cognitive Psychology*, vol. 2, pp. 145–161, 2006.

[8] M. Balakirev, *Islamey, Op. 18*. Hamburg: D. Rahter, 1902. [Online]. Available: http://imslp.org/wiki/Islamey,_Op.18_(Balakirev,_Mily)

[9] C. Sapp, "Hybrid numeric/rank similarity metrics for musical performance analysis," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2008, pp. 501–506.

[10] S. Li, D. A. A. Black, E. Chew, and M. D. Plumbley, "Evidence that phrase-level tempo variation may be represented using a limited dictionary," in *Proceedings of International Conference on Music Perception and Cognition (ICMPC'14)*, 2014.

[11] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[12] G. Claeskens and N. L. Hjort, *Model selection and Model Averaging*. Cambridge University Press, 2008.

[13] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, pp. 26 – 37, 1980.