EXPRESSIVE TIMING FROM CROSS-PERFORMANCE AND AUDIO-BASED ALIGNMENT PATTERNS: AN EXTENDED CASE STUDY

Cynthia C.S. Liem and Alan Hanjalic
Multimedia Information Retrieval Lab, Delft University of Technology, The Netherlands
{c.c.s.liem, a.hanjalic}@tudelft.nl

ABSTRACT

Audio recordings of classical music pieces reflect the artistic interpretation of the piece as seen by the recorded performing musician. With many recordings being typically available for the same music piece, multiple expressive rendition variations of this piece are obtained, many of which are induced by the underlying musical content. In earlier work, we focused on timing as a means of expressivity, and proposed a light-weight, unsupervised and audio-based method to study timing deviations among different performances through alignment patterns. By using the standard deviation of alignment patterns as a measure for the display of individuality in a recording, structural and interpretational aspects of a music piece turned out to be highlighted in a qualitative case study on five Chopin mazurkas. In this paper, we propose an entropy-based deviation measure as an alternative to the existing standard deviation measure. The obtained results for multiple short-time window resolutions, both from a quantitative and qualitative perspective, strengthen our earlier finding that the found patterns are musically informative and confirm that entropy is a good alternative measure for highlighting expressive timing deviations in recordings.

1. INTRODUCTION

In classical music, music pieces are usually conceived by composers and translated into scores. These are studied and interpreted by musicians, who each give their own personal, expressive account of the score through their actual performance of the piece. With an increasing number of such performances becoming available in digital form, we also gain access to many different artistic readings of music pieces.

The availability of recordings of multiple performances of music pieces previously has strongly been exploited in the field of audio similarity-based retrieval. In this, the focus was on matching musically closely related fragments (*audio matching* [6,8]), or finding different versions of a song at the document level, ranging from different performances of the same notated score (*opus retrieval* [2]) to potentially radically different new renditions of a previously recorded song (*cover song identification* [11]). In general, matching and retrieval of classical music pieces were shown to be achievable with near-perfect results [1, 4]. Another category of previous work largely focused on analyzing and/or visualizing the playing characteristics of individual performers in comparison to other performers [3, 9, 10].

At certain moments, a performer will display larger personal expressive freedom than at other moments, guided by theoretical and stylistic musical domain knowledge as well as personal taste and emotion. By comparing expressive manifestations in multiple recordings of the same piece, we therefore can gain insight in places in the piece where the notated musical content invites performers to display more or less expressive individualism. Such information on the interplay between performance aspects and the notated musical content provides a novel perspective on the implicit interpretative aspects of the content, which can be of a direct benefit for many Music Information Retrieval (MIR) tasks, ranging from music-historical performance school analysis to quick and informed differentiating and previewing of multiple recordings of the same piece in large databases.

In recent previous work [5], we proposed a light-weight, unsupervised and audio-based method to study timing deviations among different performances. The results of a qualitative study obtained for 5 Chopin mazurkas showed that timing individualism as inferred by our method can be related to the structure of a music piece, and even highlight interpretational aspects of a piece that are not necessarily visible from the musical score. In this paper, we introduce an entropy-based approach as an alternative to our previous standard deviation-based approach, and will study the characteristics of both methods in more depth at multiple short-time window resolutions. While this task does not have a clear-cut ground truth, the introduction of our new entropy method allows for quantitative comparative analyses, providing deeper and more generalizable insight into our meth-

ods than the largely qualitative pioneering analyses from [5].

This paper is organized as follows. After a summary of our previous work from [5], we will describe our new entropy-based method. This will be followed by a description of the experimental setup and corresponding results. Finally, the paper will end with a conclusion and discussion of future directions.

2. AUDIO-BASED ALIGNMENT AND ANALYSIS OF MULTIPLE PERFORMANCES

2.1 Audio-based alignment of multiple performances

In [5], we proposed a method to infer timing expressivity in an audio-based, objective and unsupervised data-driven way, largely building on novel work in audio similarity-based retrieval.

As short-time harmonic audio signal descriptor features, we adopt the recent Chroma Discrete Cosine Transform-reduced Log Pitch (CRP) features, which outperformed traditional chroma representations in timbre-robustness and audio matching performance [7]. We use the CRP feature implementation as made available by the original authors [1] . If A is a set with n audio recordings of the same piece, we obtain n CRP profile vectors r establishing a set R, where each r represents an audio recording $a \in A$.

As different performances of the same piece may differ in global tempo, the CRP profile vectors $r \in R$ will have different lengths. Through Dynamic Time Warping (DTW) techniques, we can align the vectors and find a time mapping between corresponding events in different recordings. For this, we apply the DTW alignment technique from [11], which used a binary cost measure and imposed local constraints to avoid pathological warpings. This method was shown to be very powerful in cover song retrieval settings. We choose a CRP profile vector $r_{ref} \in R$, corresponding to a reference recording that may be arbitrary chosen. By aligning r_{ref} with the vectors $r \in R \backslash \{r_{ref}\}$, corresponding to all other recordings in the set, full alignment between performances is achieved through r_{ref}. For each alignment between r_{ref} and an $r \in R$, an alignment matrix X is constructed. The alignment value $X_{i,j}$ between two CRP profiles at time instances i and j in r_{ref} and r, respectively ($r_{ref}[i]$ and $r[j]$), is computed adopting the local constraints as suggested in [11]. Initialization procedures, binary similarity measures and other parameters were also taken from this article, to which the interested reader is referred for more details.

An explicit alignment path is obtained by tracing back from the point corresponding to the highest total alignment score. If $|r_{ref}| = m$, for each alignment to a performance r we obtain an alignment path w of length m, with $w[1 \dots m]$ indicating short-time instance indices of the CRP profiles in r that align to $r_{ref}[1 \dots m]$. Not all time instances $1 \dots m$ may have been explicitely covered in the original alignment path. Assuming linear development for unknown instances, missing values are estimated through linear interpolation.

[1] `http://www.mpi-inf.mpg.de/~mmueller/chromatoolbox/`

2.2 Performance alignment analysis

After calculating all alignment paths following the procedures above, we will have obtained a set W with $n-1$ alignment paths $w \in W$, each of length m. We post-process these paths to emphasize irregular alignment behavior: if an alignment subpath $w[k \dots l]$ shows constant alignment steps ($w[k] = w[k+1] = w[k+2] = \dots = w[l-1] = w[l]$), this means that the corresponding CRP feature vector excerpt in r is a linearly scaled version of $r_{ref}[k \dots l]$, and therefore does not reflect any timing individualism. In order to highlight alignment step slope changes, we compute discrete second derivatives over the alignment path.

First of all, for each alignment path w, we compute the discrete first derivative δ through the central difference:

$$\delta[i] = \begin{cases} \frac{1}{2}(w[i+1] - w[i-1]) & 1 \leq i \leq m \\ w[1] - w[0] & i = 1 \\ w[m] - w[m-1] & i = m. \end{cases}$$

Due to an initial alignment index jump, a large 'startup' derivative is found at the beginning of the path. As we are only interested in the alignment step development within the true alignment path (and the beginning of the recording for the given time sampling rate will contain silence), we set the derivative values up to this startup point to 0. By repeating the central difference procedure on the enhanced δ, a second derivative approximation $\delta^2 \in \Delta^2$ is obtained.

We assume that moments in the piece showing the largest timing deviations among performers (and thus, the highest degree of individualism) must have given the performers a reason to do so, and therefore must be of a certain semantic relevance. A measure is needed to express this individuality of timing at all short-time instances of Δ^2. For this, we proposed to adopt the standard deviation: for each time instance $t = 1 \dots m$, we compute $\sigma[t]$, which is the standard deviation of all alignment second derivatives $\delta^2[t] \in \Delta^2$, acquiring a standard deviation sequence σ of length m.

3. ENTROPY AS INFORMATION MEASURE

The assumption that moments with the largest timing deviations ('disagreement') among performers will be of a certain semantic relevance resembles the notion of entropy in information theory, where items with the most uncertain actual realization are considered to hold the largest amount of information. Thus, as an alternative to our previous standard

deviation method, we now propose to calculate the entropy of Δ^2 at each short-time instance. If Δ^2 has the possible values ('symbols') $d^2_{t,1} \ldots d^2_{t,f}$ at time t, then

$$h[t] = -\sum_{i=1}^{f} p(d^2_{t,i}) \log_2 p(d^2_{t,i})$$

where we approximate $p(d^2_{t,i})$ by the frequency of $d^2_{t,i}$ in Δ^2 at time instance t. While the previous standard deviation-based approach treats the values at each $\delta^2[t]$ as cardinal data, the entropy-based approach will treat the values as nominal data, only measuring diversity.

4. EXPERIMENTAL EVALUATION

We initially conceived our methods with the goal to reveal implicitly encoded expressive musical information in audio that would go beyond an objective score reading. This means that no explicit classification is applied and an objective ground truth is absent. Because of this, in [5], the results of the standard deviation-based method were largely discussed in a qualititative way. With our new entropy-based method, possibilities arise for quantitative comparisons between this method and the standard deviation-based method, which we will discuss in this section, as an addition to qualitative and musical interpretations of the results of the entropy-based method.

Our experiments will focus on two aspects: (1) verifying that σ and h are no random noise sequences and (2) focusing on the main similarities and dissimilarities between σ and h from a quantitative and qualitative perspective. While the work in [5] only focused on a 2048-sample short-time audio analysis window, our current experiments will consider multiple possible window lengths. While we are not striving to identify an 'optimal' time window length yet (which will depend on the desired musical unit resolution, e.g. small ornamental notes vs. harmonies on beats), we consider these multiple window lengths to verify if the behavior of our methods is stable enough to not only yield interpretable results at the earlier studied resolution of 2048 samples.

4.1 Experimental Setup

Following our earlier work, we focus on 5 Chopin mazurkas that were thoroughly annotated as part of the CHARM Mazurka Project [9]: op. 17 no. 4, op. 24 no. 2, op. 30 no. 2, op. 63 no. 3 and op. 68 no. 3, with 94, 65, 60, 88 and 51 available recordings, respectively. We follow the procedure as outlined in Section 2.1, choosing the shortest recording for which manually annotated beat data is available as the reference recording, thus minimizing the size of the alignment paths. In order to interpret the results, we will use manual musical structure analyses by the authors as a reference. Thanks to the carefully established manual beat annotations

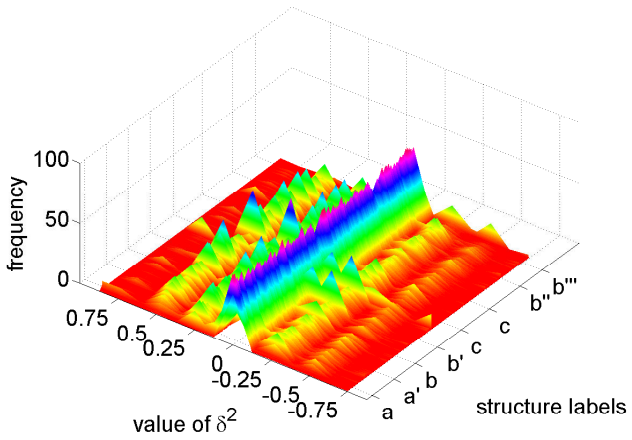


Figure 1. Histogram for δ^2 values in Δ^2 measured at consecutive short-time windows for mazurka op. 30 no. 2, for a 2048-sample window length and with reference main structural boundary labels (a, b, c, etc.) indicated over the time dimension.

from the Mazurka dataset, these structure analyses can be related to the audio as precisely as possible.

We apply our methods to all available recordings in each of the mazurkas, calculating standard deviations σ and entropies h for the alignment pattern second derivatives in Δ^2, as obtained for 7 different short-time window lengths (from 1024 to 4096 samples, in linearly increasing steps of 512 samples, at a sampling frequency of 22050 Hz and with 50% overlap). A representative example of second derivative value frequencies over the short-time instances is shown in Figure 1: the majority of values is zero ('constant alignment development'), and frequency peaks for other values appear to occur in bursts.

4.2 Verification of trends in standard deviations and entropies

To verify that both the sequences σ and h are no random noise sequences, we perform two statistical runs tests: one testing the distribution of values above and under the sequence mean, and one testing the distribution of upward and downward runs. In both cases and for all window lengths, the tests very strongly reject the null hypothesis that the sequences are random. In Figure 2, the runs frequencies for the test focusing on upward and downward runs are plotted. From this plot, we notice that entropy sequences consistently have less up- and downward runs (and thus 'smoother behavior') than standard deviation sequences, especially for small window sizes. Furthermore, the relation between the number of runs and the window size does not appear to be linear, implying that the choice of a larger short-time win-

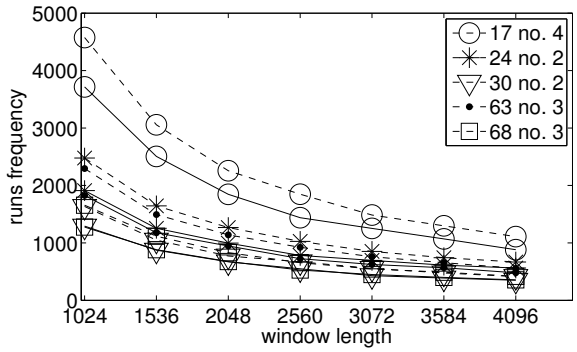


Figure 2. Numbers of up- and downward runs (summed) for different short-time window lengths. Dashed lines indicate σ sequences, solid lines indicate h sequences. Markers indicate mazurkas.

dow does not uniformly smooth the results obtained with a smaller window. Curves for the test focusing on values above and under the sequence mean are omitted due to space considerations, but strongly resemble the given plot. When plotting the resulting sequences over time, the resulting h curves indeed are less noisy than the σ curves. Figure 3 shows both curves for the opening phrase of mazurka op. 17 no. 4 for a short-time window of 1024 samples. The σ curve appears to be denser, due to the larger number of up- and downward runs. Looking at the general development of the curves, both σ and h appear to show very similar behavior, with many co-occurring maxima and minima. As a quantitative backing for this notion, Table 1 shows Pearson's correlation coefficient between σ and h for all window lengths considered. From the values in this table, it indeed becomes clear that σ and h are strongly correlated.

4.3 Standard deviations vs. entropies

As mentioned above, entropy sequences h are strongly correlated with standard deviation sequences σ. Thus, as with the σ sequences, they will be capable of highlighting developments that musically make sense [5]. Next to the example in Figure 3, where both the σ and h values increased with ornamentational variation, we also give an example where the musical score does not clearly indicate the expressive development of phrases. In Figure 4, the 'c' section of mazurka op. 30 no. 2 is shown, where a simple subphrase is almost identically repeated 8 times. A performer will not play this subphrase 8 times in an identical way, and this is reflected both in σ and h: the major displays of individuality in recordings can be found in subphrases 1 (first statement of subphrase), 3 (following traditional binary period structures, here a new subphrase could be starting, but this is not the case) and 8 (last statement of subphrase). Furthermore,

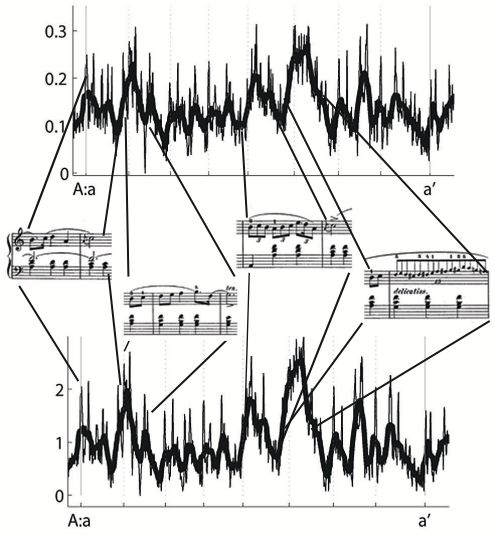


Figure 3. σ (top) and h (bottom) sequence for opening phrase of mazurka op. 17 no. 4 with corresponding score fragments. 1024-sample window length, 20-point moving average smoothed trendline indicated with thick line.

for subphrase 4 and 8, the average value of σ and h is higher than in the other subphrases, and no minima are reached as large as in the other phrases. This can be explained because of the altered ornament starting the subphrase, and the fact that both subphrase 4 and 8 are the final subphrase in a higher-order phrase hierarchy of 4 + 4 subphrases. From both Figure 3 and 4, the main difference between σ and h appears to be that h has a considerably larger range than σ, and especially tends to amplify positive peaks.

With its less noisy behavior and stronger peak amplification, the entropy-based method seems more attractive for our alignment analyses than the standard deviation-based method. As a final experiment aimed at gaining more insight into the differences between both methods, we linearly scale both σ and h to unit range. This results in sequences σ_{norm} and h_{norm}. We then test how often $h_{norm} > \sigma_{norm}$ for three cases: (1) all short-time instances, (2) all beat starts (with the beat timings obtained from the earlier manual annotations from the CHARM project) and (3) all subphrase starts. While these cases consider a decreasing number of events, the musical importance of the events increases: a subphrase start should be more informative than a random instance in time. Results are given in Table 2.

In general, σ_{norm} will have larger values than h_{norm}. This matches with the notion that the entropy sequences amplify positive peaks: thus, the non-peak values will tend to skew under the mean entropy value, while standard devia-

	1024	1536	2048	2560	3072	3584	4096
17 no. 4	0.9271	0.9225	0.9184	0.9117	0.9089	0.9022	0.9007
24 no. 2	0.9352	0.9308	0.9245	0.9218	0.9104	0.9105	0.9045
30 no. 2	0.9107	0.9094	0.9138	0.8955	0.8952	0.8911	0.8945
63 no. 3	0.9165	0.9103	0.9113	0.8992	0.8930	0.8877	0.8876
68 no. 3	0.9261	0.9274	0.9302	0.9387	0.9333	0.9291	0.9321

Table 1. Pearson's correlation coefficient between σ and h sequences for all five mazurkas with different short-time window lengths (in samples).

	1024	1536	2048	2560	3072	3584	4096
17 no. 4 overall	0.2736	0.2595	0.3994	0.3413	0.4303	0.2847	0.6966
17 no. 4 at beat starts	0.4217	0.3460	0.4798	0.3662	0.4571	0.2955	0.7020
17 no. 4 at subphrase starts	0.6462	0.5077	0.6769	0.4769	0.5231	0.4462	0.7385
24 no. 2 overall	0.3645	0.5912	0.3172	0.4754	0.6417	0.5548	0.7307
24 no. 2 at beat starts	0.4903	0.6842	0.3767	0.5097	0.6898	0.5845	0.7895
24 no. 2 at subphrase starts	0.5085	0.7288	0.3559	0.5254	0.7966	0.6271	0.8644
30 no. 2 overall	0.2238	0.2354	0.1944	0.1790	0.3030	0.4177	0.6508
30 no. 2 at beat starts	0.3212	0.3005	0.1606	0.1762	0.2902	0.4301	0.6321
30 no. 2 at subphrase starts	0.4375	0.4375	0.3125	0.3438	0.3750	0.5000	0.8125
63 no. 3 overall	0.4901	0.5869	0.7861	0.6578	0.8038	0.5617	0.5956
63 no. 3 at beat starts	0.6348	0.6565	0.8348	0.6696	0.8261	0.5435	0.5739
63 no. 3 at subphrase starts	0.8684	0.8947	0.9474	0.7895	0.8421	0.5789	0.6053
68 no. 3 overall	0.1574	0.3359	0.1383	0.2698	0.6095	0.4751	0.6628
68 no. 3 at beat starts	0.3039	0.4420	0.1823	0.3094	0.6575	0.5304	0.6906
68 no. 3 at subphrase starts	0.3000	0.5000	0.2333	0.4000	0.6333	0.7000	0.7000

Table 2. Normalized entropies h_{norm} vs. standard deviations σ_{norm}: fractions of cases in which $h_{norm} > \sigma_{norm}$ considered over all short-time instances, over all beat starts, and over all subphrase starts different short-time window lengths (in samples).

tions are centered around the mean in a more balanced way. Mazurka op. 63 no. 3 is an exception, but this may have been caused by the noisiness of the historical reference recording (Niedzielski 1931), which causes clicking and hissing effects at random moments throughout the piece, thus also causing irregular alignment behavior at these random moments. However, in all cases, when only looking at time instances with beat and subphrase starts, the fraction of larger normalized entropies increases for all mazurkas. Especially for subphrases in comparison to beat starts, the increase is considerable. This implies that the entropy sequence values indeed amplify musically meaningful peaks.

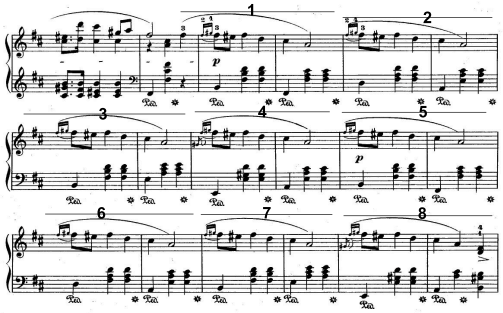
Looking at the differences between beat start and subphrase start fractions, the increases initially may not appear to be stable or generalizable over different mazurkas. For subphrase starts, the probability that $h_{norm} > \sigma_{norm}$ is much larger than for beat starts in mazurkas op. 17 no. 4 and op. 63 no. 3 (and to a lesser extent, op. 30 no. 2). On the other hand, in mazurkas op. 24 no. 2 and op. 68 no. 3, this is much less the case, with the beat and subphrase start fractions being much closer to each other.

From a musical perspective, this may not seem as strange as from a numerical perspective: mazurkas op. 24 no. 2 and op. 68 no. 3 both are rather 'straightforward' pieces, with many repeating blocks with little thematic development, and constant ongoing rhythms. Thus, there is not so much flexibility to shape structural boundaries and subphrase starts with large timing differences. On the other hand, mazurkas op. 17 no. 4 and op. 63 no. 3 are very dramatical, have strongly differing thematic blocks, and thus allow for emphasizing of new subphrases. While resembling mazurkas op. 24 no. 2 and op. 68 no. 3 in terms of rhythmical and thematic straightforwardness, mazurka op. 30 no. 2 is less rigid in terms of phrasing and musical movement, and thus will allow for more timing flexibility, thus also sharing characteristics with the other two mazurkas.

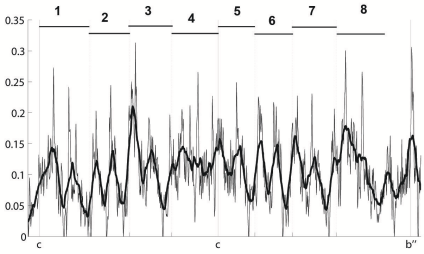
5. CONCLUSION AND RECOMMENDATIONS

In this paper, we proposed an entropy-based method as an alternative to a standard deviation-based method for studying alignment patterns between multiple audio recordings, which were considered to contain interesting information about the recorded music that cannot objectively be inferred from a score. Our entropy method yielded results that consistently were strongly correlated with the standard deviation results at multiple time resolutions, while being less noisy and amplifying positive peaks, which both are desirable properties for our purposes. It was shown that both the standard deviation and entropy methods do not depict random noise, but can be related to actual musical content.

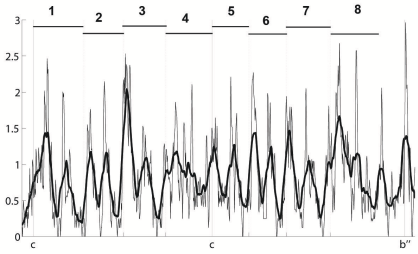
The development over multiple time resolutions of correlations between standard deviation and entropy sequences, the frequencies of up- and downward runs, as well as runs



(a) Score with numbered subphrases



(b) Standard deviation sequence σ



(c) Entropy sequence h

Figure 4. Mazurka op. 30 no. 2, σ and h for 'c' section. The 8 repeating subphrases are numbered. 1024-sample window length, 20-point moving average smoothed trendline.

above and under the sequence mean, yields similar trends over different mazurkas, implying that our methods are generalizable. We did not focus yet on further implications of the choice of short-time window length, which still needs to be done in future work. Another main future challenge is the further solidification and backing of the musical interpretations of our results. Finally, we did not yet employ any noise-filtering or signal enhancement techniques. While the results obtained for the noisy op. 68 no. 3 Niedzielski reference recording on runs frequency and correlation trends are largely consistent with the results for other mazurkas with clean reference recordings, the reference recording quality will influence results and this topic should be investigated more in future work.

Rendering MIDI files as audio and modifying them in a controlled way may partially overcome the problem of a missing ground truth and possible noise in real-life reference recordings. In addition, the interpretation of results can be strengthened through a combination of our methods with other MIR techniques dealing with prior knowledge of the musical content in a more explicit and supervised way. Supported by our methods, such techniques will not have to be tediously applied to a full database, but can be limited to one or more reference recordings. This introduces promising directions for MIR tasks dealing with the real-life abundance of artistically valuable digital recordings.

Acknowledgements: Cynthia Liem is a recipient of the Google European Doctoral Fellowship in Multimedia, and this research is supported in part by this Google Fellowship.

6. REFERENCES

[1] M. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Trans. on Audio, Speech and Language Proc.*, 16(5):1015–1028, July 2008.

[2] M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proc. of the IEEE*, 96(4):668–696, April 2008.

[3] M. Grachten and G. Widmer. Who is who in the end? Recognizing pianists by their final ritardandi. In *Proc. Intl. Soc. for MIR Conf. (ISMIR)*, Kobe, Japan, October 2009.

[4] C.C.S. Liem and A. Hanjalic. Cover song retrieval: A comparative study of system component choices. In *Proc. Intl. Soc. for MIR Conf. (ISMIR)*, Kobe, Japan, October 2009.

[5] C.C.S. Liem, A. Hanjalic, and C.S. Sapp. Expressivity in musical timing in relation to musical structure and interpretation: A cross-performance, audio-based approach. In *Proc. 42nd Int. AES Conf. on Semantic Audio*, pages 255–264, Ilmenau, Germany, July 2011.

[6] M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[7] M. Müller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE Trans. on Audio, Speech and Language Proc.*, 18:649–662, March 2010.

[8] M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proc. Intl. Conf. on MIR (ISMIR)*, pages 288–295, 2005.

[9] C.S. Sapp. Comparative analysis of multiple musical performances. In *Proc. Intl. Conf. on MIR (ISMIR)*, Vienna, Austria, September 2007.

[10] C.S. Sapp. Hybrid numeric/rank similarity metrics for musical performance analysis. In *Proc. Intl. Conf. on MIR (ISMIR)*, Philadelphia, USA, September 2008.

[11] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Trans. on Audio, Speech and Language Proc.*, 16:1138–1151, August 2008.