Framework for Exploration of Performance Space

Özgür İzmirli
Department of Computer Science
Connecticut College, CT, USA
oizm@conncoll.edu

ABSTRACT

This paper presents a framework for the analysis and exploration of performance space. It enables the user to visualize performances in relation to other performances of the same piece based on a set of features extracted from audio. A performance space is formed from a set of performances through spectral analysis, alignment, dimensionality reduction and visualization. Operation of the system is demonstrated initially with synthetic MIDI performances and then with a case study of recorded piano performances.

Author Keywords

performance space, visualization, exploration framework

1. INTRODUCTION

The music score is a representation which contains information about the composer's intention. In a classical setting the performer interprets the score and prepares the performance by rehearsing the piece with his/her own understanding guided by the sense of aesthetics. While the audience hears one outcome of this exploration at the concert, the performer spends a considerable amount of time experimenting with various subtleties in phrasing, articulation, tone and projection. Performers play an active role in shaping performance and their performances reflect the result of their exploration (e.g. [11]). Pitch, duration, rhythm, dynamics, accents and other performance parameters are usually indicated in the score but these fail to capture sonic and time related performance details such as expressive timing and stylistic phrasing, thus forming a gap between the score and its performance. The information that resides within the score-performance gap is by no means redundant and makes for a large part of the musical experience. In today's practice we can capture performances through audio, or to an extent, symbolic (MIDI) recordings. However, once recorded these performances are stored linearly, identified through metadata and therefore largely remain opaque to exploration. The presented system aims at utilizing the information within this gap for purposes of analysis, cataloging and exploration.

In this paper, we propose a flexible and modular framework for visualizing performance space. Expressive performance has many attributes and some of these are easily quantifiable while others are not and remain as open problems. In this work we focus on tempo and dynamics as the two most important attributes of expressive performance. In the remainder of the paper we give a brief list of related work and then describe the proposed framework. We then provide examples incrementally to explain and demonstrate the operation of one particular realization of the framework. We provide examples from synthesized and actual performances.

NIME'15, May 31-June 3, 2015, Louisiana State Univ., Baton Rouge, LA.

2. RELATED WORK

We know that expert performers deviate significantly from the norm and strive for individuality. This section very briefly summarizes related work in expressive performance modeling and analysis. In their work analyzing Chopin's 24-2 Mazurka performances, Rink et al. [15] have pointed out that many musical gestures are weakly captured by musical notation, they are not necessarily correlated with the harmonic and rhythmic patterns and these gestures are realized through the agency of performance. Widmer et al. [19][6][4] have published extensively on expressive music performance. They proposed the performance worm which is a graphical representation showing the trajectory of the performance point in the dynamics-tempo plane which could be used to characterize performers. Sapp's scape plots [16] are visualizations of comparative performances with a range of time spans. Wang [18] devised a method for quantifying performer styles using recordings. He calculated features from performances and compared them across performers. He noted the similarities in performances between various performers.

Chew [3] argues that engineering tools are useful for visualizing musical parameters and these tools can shed light on various aspects such as composition, music cognition and performance. Fairly accurate performance analysis could be done using transcription if it were reliable. Grosche et al. [7] describe the difficulty of extracting tempo and beat information from music recordings and propose a mid-level representation that captures musically meaningful local pulse information. Repp published a number of papers on expressive timing from a music perception standpoint [14]. The Mazurka Project [1] collected recordings of Chopin's Mazurkas by different performers. Sapp [16] provided beat annotations for several of the Mazurkas for multiple performers. Unaligned MIDI files can be found separately. Sapp studied performance differences in this collection and proposed a numerical method for examining similarities among tempo and loudness features extracted from recordings [17].

3. FRAMEWORK

We define a performance space as a low-dimensional visualization that shows distance relationships of features extracted from a set of performances. The assumption is that the set consists of the same musical material, that is, the corresponding score is the same, and the variation from one piece to the other is due to the expressive performance choices. The framework for obtaining a performance space is general and the individual components can be conceptualized and implemented in different ways. Here, we present the components

of the framework and describe their specific realizations. The framework consists of the following components: spectral analysis front-end, alignment, dimensionality reduction and visualization. The choice of distances as well as the design of features play an important role in this process. Distances between two performances can be defined in many different ways depending on the degree information that can be reliably extracted from the audio of these performances. The pairwise distances are used to form a low-dimensional spatial representation that allows the user to visualize the 'commonness' or 'novelty' of a particular performance in relation to other performances in the set of pieces under consideration.

In this work we chose to use time warping costs and similarities between dynamics. We use a single musical fragment of piano music as our working example. Piano is chosen as the instrument because of its clear onsets and dynamics as well as the availability of the performance data. We have chosen to use measures 36-44 of Chopin's Mazurka 17-4 during which different performers exhibit ample variations in tempo and dynamics.

3.1 Spectral Representation

Among the many options for a time-frequency representation such as the Fourier Transform, wavelets, chroma pitch, semigram or the constant Q transform, we chose to use chroma pitch features described in [13]. The chroma pitches are computed from the input signal decomposed into 88 bands (A0 to C8) using a constant Q multirate filter bank. A filter output is calculated for each time frame. Each filter output measures the local energy content (short-time mean-square power) in its subband. The logarithmically spaced output of this filter bank provides a compact way to capture pitched content while keeping summarization of the spectrum to a minimum in contrast to, for example, the chromagram [2] which usually folds the frequency over octaves. The traditional Fourier Transform on the other hand is linear in frequency and has unnecessary resolution at the high end of the spectrum for our purposes. We use a sampling frequency of 22050 Hz and a window length of 40 ms with 50 percent overlap.

3.2 Audio Alignment and Warp Cost

Audio-to-audio and audio-to-score alignment has received ample attention in the literature and many approaches for this problem exist (e.g. [12][5][9][8]). The purpose of alignment is to find correspondences between two renditions of the same piece and align them on a frame by frame basis, without resorting to any annotations, and solely looking at their audio content. Audio-to-audio methods aim to find an optimal warping path to match spectral representations on both sides as best as possible. The audio-to-score problem is usually translated back into the audio-to-audio problem by synthesizing the symbolic score. This works reasonably well even with simplistic sound synthesis.

The alignment component is implemented using a basic version of Dynamic Time Warping (DTW) on the chroma pitch features. Cosine distance is used to calculate the similarity matrix prior to running the DTW algorithm. The DTW results in a warp path for every pair of pieces in the input sound set. The warp path reveals the time differences between corresponding musical events due to changes in instantaneous tempo. The warp path is a straight line on the diagonal of the cost matrix if the two pieces are identical. The warp path will deviate from the diagonal when we need to progress faster along one piece compared to the other.

For this module we have implemented two distance measures. One measures how similar the two performances are in terms of their relative tempo curves. Two pieces are considered similar if they tend to accelerate and decelerate during the same parts of the music. The comparison is based on relative tempo changes since we do not want absolute tempi to be a factor. The warp cost is calculated by fitting a spline curve on the downsampled warp path and averaging the differences between the spline curve and the diagonal. A distance close to zero will be obtained when the two pieces are almost identical. The farther the warp path moves away from the diagonal in either direction the higher the distance. The second measure approximates the similarity of dynamics between two performances by comparing only the energies along the warped spectra. For each frame, the wide band energy is obtained by summing energies in all bins. Then the frame sequence is smoothed with a Gaussian window over time to make it tolerant to small alignment errors. The distance is modeled with correlation distance between the two wide band energy signals.

The distance matrix D is formed by calculating all pairwise distances. We can choose to use only relative tempo, $D^t_{i,j}$ between piece i and j, only dynamics, $D^d_{i,j}$ or a weighted combination of the two, $\alpha D^t_{i,j} + (1-\alpha)D^d_{i,j}$. Both distance matrices are normalized by dividing by their maximum element. We will use the combined metric and set α to 0 or 1 when only one distance is required. We demonstrate each of these cases below.

3.3 Visualization

As is the case for most dimensionality reduction methods, Multidimensional Scaling (MDS) [10] aims to find a mapping from a high dimensional representation to a low dimensional one such that the between-element distances are preserved as much as possible. In this paper we utilize MDS to obtain 2-dimensional spaces for visualization purposes, however, theoretically this is not a limitation and the output dimension could be chosen to be higher. We use Sammon's nonlinear mapping as the goodness of fit criterion for MDS. The distance matrix D is the only input to MDS besides the output dimension.

4. PERFORMANCE SPACES

Before presenting some examples we summarize the sequence of operations carried out by the components described above. In order to construct and visualize a performance space for a given collection, initially the chroma pitch features are calculated for each performance. Next, the pairwise distance matrices are calculated for all types of distances to obtain the combined distance matrix (by setting weights). MDS is applied to the distance matrix for the given output dimension. Finally, the coordinates of the objects are plotted together either with their labels or tempo and dynamics curves.

4.1 Relative Tempo Curves

We first show how the proposed model can be used for visualization of performance space using similarity of tempo curves. In order to understand how the system maps the input sounds to a 2D output space we created parametrized tempo curves. Each tempo curve is defined by 3 points over time (beginning, middle and end of fragment). These points are generated randomly for each piece and spline interpolated for evaluation at any arbitrary time in the fragment. A value of 1 on this curve means original tempo and 1.15 means 15 percent faster tempo. A modified MIDI sequence is calculated according to the interpolated tempo curve from an original MIDI sequence that contains the music with constant tempo and flat dynamics. The audio is synthesized using the modified MIDI sequence and stored

in the collection. The performance space is calculated with $\alpha = 1$.

The output for 30 parametrized performances are shown in the left plot in Figure 1. The points are locations of the data points in the MDS output space. To the right of each point the corresponding tempo curve is shown. The horizontal line depicts unity tempo and a tempo curve on this horizontal line would result in the original MIDI timing without any tempo alterations. As far as vertical scale, in this example, the peaks represent a 15 percent swing around the unity tempo curve. The color is cycled for purposes of clarity in visualization to distinguish between overlapping curves and does not provide any additional information. It can be seen from the figure that the not so exciting performances that have relatively little tempo changes are concentrated around the middle of the space. Decreasing tempo curves have grouped in the top middle while the increasing tempo curves are in the bottom middle. To the right are U shaped curves and to the left are the inverse U shaped curves. The relationship between location in the output and the tempo contour can be clearly seen. Since the output is continuous and not categorical, a tempo curve with any shape will find a place in this 2D space.

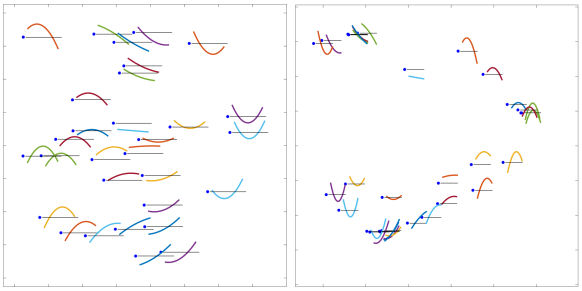

Figure 1: Output space showing relative tempo curves (left) and relative dynamics curves (right).

4.2 Relative Dynamics Curves

Similar to the relative tempo curves, we now turn to dynamics by replacing the tempo curve with a dynamics curve. The dynamics curve is found by correlating the smoothed wide band energy signals of the aligned spectra. The use of correlation distance to calculate the distance between the dynamics curves allows the dynamics comparison to be relative. That is, changes in dynamics are compared regardless of the absolute loudness. The right plot in Figure 1 shows the output for the same 30 performances for randomly generated dynamics curves ($\alpha = 0$). The contents of the figure have the same interpretation with the following exception: the range of the dynamics curves is +-57 MIDI velocity increments relative to the current velocity of the note. We use a baseline of velocity 70 in the flat MIDI file to allow for a swing in both directions. It can be clearly observed in the figure that a similar distribution of points has taken place.

4.3 Combined Dynamics and Tempo

In actual performance, tempo and dynamics changes are engaged simultaneously. We therefore provide a combined example with $\alpha = 0.5$. Figure 2 shows the output in which the thick (orange) lines are tempo curves and the thin (blue) ones are dynamics curves. Here the results are somewhat mixed possibly due to having equal weight between tempo and dynamics. Similar to the previous examples, the regions with similar patterns can be clearly identified.

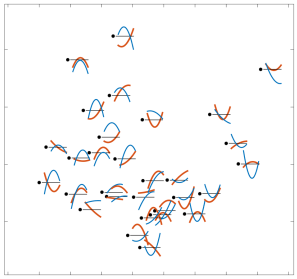

Figure 2: Output space using combined tempo and dynamics curves.

4.4 Actual Performances

Having understood the capabilities of the system we can now move on to actual performances. Figure 3 shows the output for 50 performances from actual commercial recordings with $\alpha = 0.5$. Here the data points are labeled with performer names - duplicate performer names appear because the collection contains multiple recordings by the same performer. The space reveals interesting similarities as well as differences between well-known performers.

The numbered points represent a pianist's performances specifically for testing our system. The pianist was asked to practice the fragment and then he listened to performances of the same fragment by Kissin, Kushner and Ashkenazy twice each. After listening to each performer, the pianist played the piece three times aiming for a similar interpretation but not necessarily an exact imitation as he did not have time to dissect and memorize the interpretations. Performances labeled 51-53 were played on a weighted-key velocity-sensitive electronic keyboard with sampled sounds and without a pedal after listening to Kissin. Points labeled 54-56 were performed on a baby grand after listening to Kushner. The final three labeled 57-59 were again played on a baby grand after listening to Ashkenazy. Only one performance of the first group got close to Kissin. This is probably because of the keyboard's response characteristics. Members of the second group were close to Kushner and the performances can be considered to be fairly consistent (close). The last group had the highest consistency although they did not end up being closest to Ashkenazy but in the same area.

5. DISCUSSION

Performance space is a multi-faceted concept and its models need to offer fine-grained and accurate processing algorithms that are sensitive to subtle variations. Micro timing, articulation, tone and stylistic touch are among these currently elusive attributes. In this work, we have chosen to provide realizations of the components in order to present a proof of concept. Hence we did not try to employ sophisticated methods. The modular nature of the system allows for both specialization and refinement on a component basis. The application to piano music is discussed in this paper but in another context the same framework could be used for exploring idiomatic performance details for other instruments. For each application appropriate features need to be chosen, signal processing algorithms implemented and associated distances need to be defined.

Both distances described in the paper are relative measures. We look at the changes rather than absolute tempi and dynamics. While the nature of the changes carry important information about expressive performance it is true that two performances will be received as entirely different if they are differing widely in dynamics and/or tempo. For

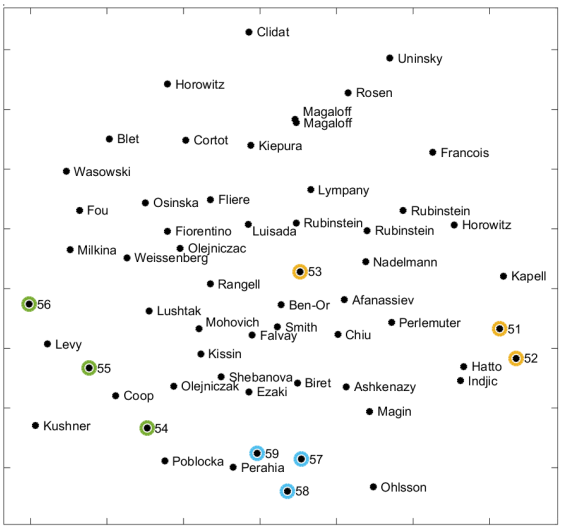

Figure 3: Output space from actual commercial recordings plus pianist performances.

example, two performances that have very different starting tempi with the same percentage speed-up throughout the piece will be heard as significantly different, but, the current system will not differentiate between the two. We assume that the performances are guided by a score or listening experiences such that the interpretations roughly aim for the composer's intended tempo and dynamics.

In the last section, our example aimed to demonstrate the use of the system as a pedagogical and personal development tool. It showed how pianists could systematically examine their performance, explore variations relative to others and even find a niche by experimentation. If, on the other hand, the pianist chose to start only with his/her own performances he would be visualizing the space within the realm of his imagination and aesthetic choices. Obviously, this exploration can be conducted iteratively such that the visualization can be obtained after each new performance. In the same vein it could be used for checking the consistency of expressive performance across multiple performances separated by considerable time.

Finally, it is interesting to note that the Hatto hoax surfaces again in this work, even in the short fragment that we are using. It has been reported in many works that recordings by Indjic were used in place of Hatto's.

6. CONCLUSION

We have presented a framework for exploration of expressive performances through visualization. We have provided realizations of the components of the framework in order to show its operation. While we recognize that many other attributes play important roles in music expression we have concentrated on tempo and dynamics in this work. Nevertheless, this is not a limitation of the framework and the system can be designed to 'hear' musically relevant aspects of performances with more sophisticated features. The output space in 2D has been shown to be quite informative and to be able to associate regions with certain shapes of tempo and dynamics curves.

7. REFERENCES

[1] The mazurka project, http://www.mazurka.org.uk.

[2] M. A. Bartsch and G. H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001.

[3] E. Chew. About time: Strategies of performance revealed in graphs. *Visions of Research in Music Education 20(1)*, 20(1):401–409, 2012.

[4] S. Dixon, W. Goebl, and G. Widmer. The performance worm: Real time visualisation of expression based on Langner's tempo-loudness animation. In *Proceedings of the international computer music conference*, 2002.

[5] S. Ewert, M. Muller, and P. Grosche. High resolution audio synchronization using chroma onset features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1869–1872, 2009.

[6] M. Grachten, W. Goebl, S. Flossmann, and G. Widmer. Phase-plane representation and visualization of gestural structure in expressive timing. *Journal of New Music Research*, 38(2):183–195, 2009.

[7] P. Grosche and M. Muller. Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1688–1701, 2011.

[8] N. Hu, R. B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[9] C. Joder, S. Essid, and G. Richard. Learning optimal features for polyphonic audio-to-score alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2118–2128, 2013.

[10] J. B. Kruskal and M. Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[11] M. Molina-Solana and M. Grachten. Nature versus culture in ritardando performances. In *Proc. Sixth Conference on Interdisciplinary Musicology*, 2010.

[12] N. Montecchio and A. Cont. A unified approach to real time audio-to-score and audio-to-audio alignment using sequential montecarlo inference techniques. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 193–196, 2011.

[13] M. Müller and S. Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the 12th International Conference on Music Information Retrieval*, Miami, USA, 2011.

[14] B. H. Repp. Variations on a theme by chopin: Relations between perception and production of timing in music. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3):791, 1998.

[15] J. Rink, N. Spiro, and N. Gold. Motive, gesture, and the analysis of performance. *New Perspectives on Music and Gesture*, pages 267–292, 2011.

[16] C. S. Sapp. Comparative analysis of multiple musical performances. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 497–500, 2007.

[17] C. S. Sapp. Hybrid numeric/rank similarity metrics for musical performance analysis. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 501–506, Philadelphia, USA, 2008.

[18] C.-i. Wang. Quantifying pianist style-an investigation of performer space and expressive gestures from audio recordings. Master's thesis, New York University, 2013.

[19] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic. In search of the Horowitz factor. *AI Magazine*, 24(3):111, 2003.