HYBRID NUMERIC/RANK SIMILARITY METRICS FOR MUSICAL PERFORMANCE ANALYSIS

Craig Stuart Sapp
CHARM, Royal Holloway, University of London
craig.sapp@rhul.ac.uk

ABSTRACT

This paper describes a numerical method for examining similarities among tempo and loudness features extracted from recordings of the same musical work and evaluates its effectiveness compared to Pearson correlation. Starting with correlation at multiple timescales, other concepts such as a performance "noise-floor" are used to generate measurements which are more refined than correlation alone. The measurements are evaluated and compared to plain correlation in their ability to identify performances of the same Chopin mazurka played by the same pianist out of a collection of recordings by various pianists.

1 INTRODUCTION

As part of the Mazurka Project at the AHRC Centre for the History and Analysis of Recorded Music (CHARM), almost 3,000 recordings of Chopin mazurkas were collected to analyze the stylistic evolution of piano playing over the past 100 years of recording history, which equates to about 60 performances of each mazurka. The earliest collected performance was recorded on wax cylinders in 1902 and the most recent posted as homemade videos on YouTube. Table 1 lists 300 performances of five mazurkas which will be used for evaluation later in this paper since they include a substantial number of recordings with extracted tempo and loudness features.

Mazurka		*Performances*	
Opus	*Key*	*Collected*	*Processed*
17/4	A minor	93	63
24/2	C major	63	29
30/2	B minor	60	33
63/3	C♯ minor	87	57
68/3	F major	51	50

Table 1. Collection of musical works used for analysis.

For each of the processed recordings, beat timings in the performance are determined using the *Sonic Visualiser* audio editor [1] for markup and manual correction with the assistance of several vamp plugins. [2] Dynamics are then extracted as smoothed loudness values sampled at the beat positions.[3] Feature data will eventually be extracted from all collected mazurkas in the above list, but comparisons made in Section 3 are based on the processed performance counts in Table 1. Raw data used for analysis in this paper is available on the web. [3]

Figure 1 illustrates extracted performance feature data as a set of curves. Curve 1a plots the beat-level tempo which is calculated from the duration between adjacent beat timings in the recording. For analysis comparisons, the tempo curve is also split into high- and low-frequency components with linear filtering. [4] Curve 1b represents smoothed tempo which captures large-scale phrasing architecture in the performance (note there are eight phrases in this example). Curve 1c represents the difference between Curves 1a and 1b which is called here the desmoothed tempo curve, or the residual tempo. This high-frequency tempo component encodes temporal accentuation in the music used by the performer to emphasize particular notes or beats. Mazurka performances contain significant high-frequency tempo information, since part of the performance style depends on a non-uniform tempo throughout the measure—the first beat usually shortened, while the second and/or third beat are lengthened. Curve 1d represents the extracted dynamics curve which is a sampling of the audio loudness at each beat location.

Other musical features are currently ignored here, yet are important in characterizing a performance. In particular, pianists do not always play left- and right-hand notes together, according to aural traditions, although they are written as simultaneities in the printed score. Articulations such as legato and staccato are also important performance features but are equally difficult to extract reliably from audio data. Nonetheless, tempo and dynamic features are useful for developing navigational tools which allow listeners to focus their attention on specific areas for further analysis.

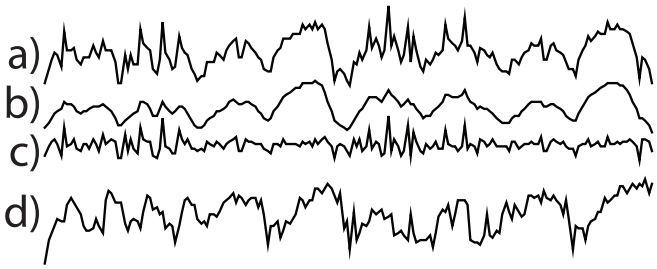

Figure 1. Extracted musical features from a recording of Chopin's mazurka in B minor, 30/2: a) tempo between beats; b) smoothed tempo; c) residual tempo ($c = a - b$); and d) beat-level dynamics.

[1] http://www.sonicvisualiser.org
[2] http://sv.mazurka.org.uk/download
[3] http://mazurka.org.uk/info/excel
[4] The filtering method is available online at http://mazurka.org.uk/software/online/smoother .

2 DERIVATIONS AND DEFINITIONS

Starting with the underlying comparison method of correlation (called S_0 below), a series of intermediate similarity measurements (S_1, S_2, and S_3) are used to derive a final measurement technique (S_4). Section 3 then compares the effectiveness of S_0 and S_4 measurements in identifying recordings of the same performer out of a database of recordings of the same mazurka.

2.1 Type-0 Score

As a starting point for comparison between performance features, Pearson correlation, often called an r-value in statistics, is used:

$$\text{Pearson}(x, y) = \frac{\sum_n (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_n (x_n - \bar{x})^2 \sum_n (y_n - \bar{y})^2}} \quad (1)$$

This type of correlation is related to dot-product correlation used in Fourier analysis, for example, to measure similarities between an audio signal and a set of harmonically related sinusoids. The value range for Pearson correlation is -1.0 to $+1.0$, with 1.0 indicating an identical match between two sequences (exclusive of scaling and shifting), and the value 0.0 indicating no predictable linear relation between the two sequences x and y.

Correlation values between extracted musical features typically have a range between 0.20 and 0.97 for different performances of mazurkas. Figure 2 illustrates the range of correlations between performances in two mazurkas. Mazurka 17/4 is a more complex composition with a more varied interpretation range, so the *mode*, or most-expected value, of the correlation distribution is 0.67. Mazurka 68/3 is a simpler composition with fewer options for individual interpretations so the mode is much higher at 0.87.

These differences in expected correlation values between two randomly selected performances illustrate a difficulty in interpreting similarity directly from correlation values. The correlation values are consistent only in relation to a particular composition, and these absolute values cannot be compared directly between different mazurkas. For example, a pair of performances which correlate at 0.80 in mazurka 17/4 indicates a better than average match, while the same correlation value in mazurka 68/3 would be a relatively poor match. In addition, correlations at different timescales in the same piece will have a similar problem, since some regions of music may allow for a freer interpretation while other regions may have a more static interpretation.

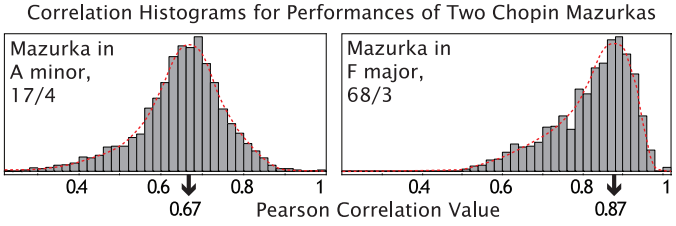

Figure 2. Different compositions will have different expected correlation distributions between performances.

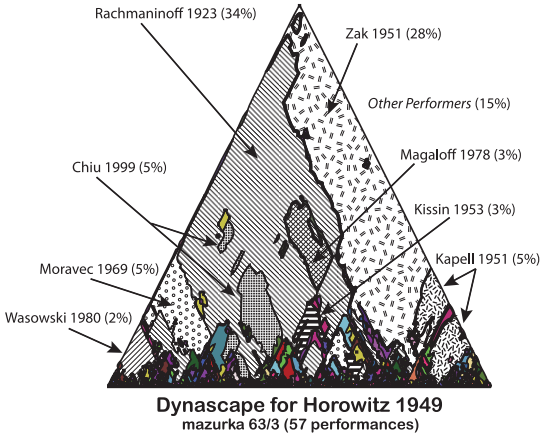

Figure 3. Scapeplot for the dynamics in Horowitz's 1949 performance of mazurka 63/3 with the top eight matching performances labeled.

2.2 Type-1 Score

In order to compensate partially for this variability in correlation distributions, scapeplots were developed which only display nearest-neighbor performances in terms of correlation at all timescales for a particular reference performance.[3] Examples of such plots created for the Mazurka Project can be viewed online. [5]

The S_1 score is defined as the fraction of area each target performance covers in a reference performer's scapeplot. Figure 3 demonstrates one of these plots, comparing the dynamics of a performance by Vladimir Horowitz to 56 other recordings. In this case, Rachmaninoff's performance of the same piece matches better than any other performance, since his performance covers 34% of the scape's plotting domain. At second best, Zak's performance matches well towards the end of the music, but covers only 28% of the total plotting domain. Note that Zak's performance has the best correlation for the entire feature sequence (S_0 score) which is represented by the point at the top of the triangle. The S_1 scores for the top eight matches in Figure 3 are listed in Table 2. There is general agreement between S_1 and S_0 scores since five top-level correlation matches also appear in the list.

2.3 Type-2 Score

Scape displays are sensitive to the *Hatto effect*: if an identical performance to the reference, or query, performance is present in the target set of performances, then correlation values at all time resolutions will be close to the maximum value for the identical performance, and the comparative scapeplot will show a solid color. All other performances would have an S_1 score of approximately 0 regardless of how similar they might otherwise seem to the reference performance. This property of S_1 scores is useful for identifying two identical recordings, but not useful for viewing similarities to other performances which are hidden behind such closely neighboring performances.

One way to compensate for this problem is to remove the best match from the scape plot in order to calculate the next best match. For example, Figure 4 gives a rough schematic for

[5] http://mazurka.org.uk/ana/pcor-perf

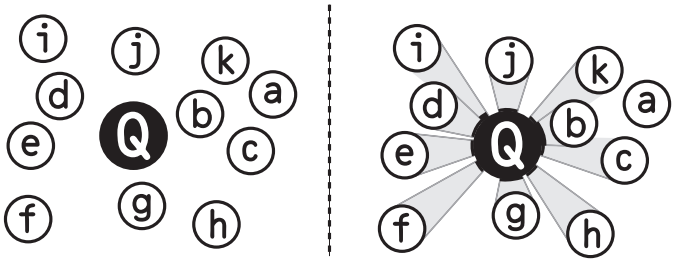

Figure 4. Schematic of nearest-neighbor matching method used in comparative timescapes.

how scapeplots are generated. Capital 'Q' represents the query, or reference, performance, and lower-case lettered points represent other performances. The scapeplot basically looks around in the local neighborhood of the feature space and displays closest matches as indicated by lines drawn towards the query on the right side of Figure 4. Closer performances will tend to have larger shadows on the query, and some performances can be completely blocked by others, as is the case for point "*a*" in the illustration.

An S_2 score measures the coverage area of the most dominant performance which is assumed to be most similar to the reference performance. This nearest of the neighbors is then removed from the search database, and a new scapeplot is generated with the remaining performances. Gradually, more and more performances will be removed which allows for previously hidden performances to appear in the plot. For example, point "*a*" in Figure 4 will start to become visible once point "*b*" is removed. S_2 scores and ranks are independent. As fewer and fewer target matches remain, the S_2 scores will increase towards 1.0 while the S_2 ranks decrease towards the bottom match.

In Figure 3, Rachmaninoff is initially the best match in terms of S_1 scores, so his performance will be removed and a new scapeplot generated. When this is done, Zak's performance then represents the best match, covering 34% of the scapeplot. Zak's performance is then removed, a new scapeplot is calculated, and Moravec's performance will have the best coverage at 13%, and so on. Some of the top S_2 scores for Horowitz's performance are listed in Table 2.

2.4 Type-3 Score

Continuing on, the next best S_2 rank is for Chiu, who has 20% coverage. Notice that this is greater than Moravec's score of 13%. This demonstrates the occurrence of what might be called the lesser Hatto effect: much of Moravec's performance overlapped onto Chiu's region, so when looking only at the nearest neighbors in this manner, there are still overlap problems. Undoubtedly, Rachmaninoff's and Zak's performances mutually overlap each other in Figure 3 as well. Both of them are good matches to Horowitz, so it is difficult to determine accurately which performance matches "best" according to S_2 scores since they are interfering with each others scores and are both tied at 34% coverage.

In order to define a more equitable similarity metric and remove the Hatto effect completely, all performances are first ranked approximately by similarity to Horowitz using either

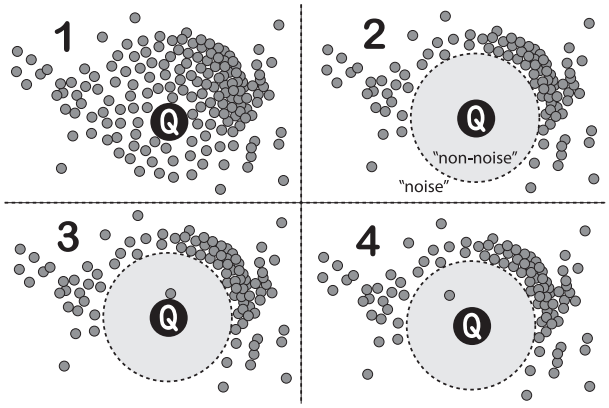

Figure 5. Schematic of the steps for measuring an S_3 score: (1) sort performances from similar to dissimilar, (2) remove most similar performance to leave noise floor, (3) & (4) insert more similar performances one-by-one to observe how well they can occlude the noise-floor.

S_0 values or the rankings produced during S_2 score calculations. Performances are then divided into two groups, with the poorly-matching half defined as the performance "noise-floor" over which better matches will be individually placed.

To generate an S_3 score, the non-noise performances are removed from the search database as illustrated in step 2 of Figure 5, leaving only background-noise performances. Next, non-noise performances are re-introduced separately along with all of the noise-floor performances and scapeplot is generated. The coverage area of the single non-noise performance represented in the plot is defined as its S_3 similarity measurement with respect to the query performance.

This definition of a performance noise-floor is somewhat ar-

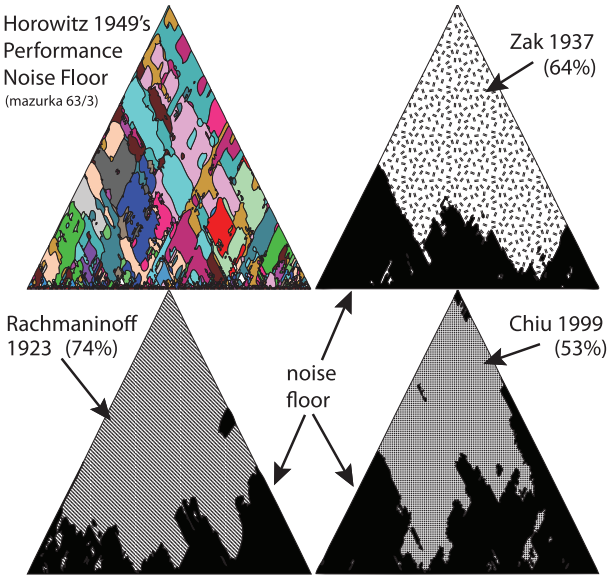

Figure 6. Dynascapes for Horowitz's performance of mazurka 63/3. Top left is a plot of the noise-floor performances, and the other three plots separate include one of the top matching performances which can cover most of the noise-floor.

Target	S_0	R_0	S_1	S_2	S_3	S_{3r}	S_4	R_4
Rac23	0.60	3	0.34	0.34	0.74	0.82	0.78	1
Zak37	0.64	1	0.28	0.34	0.64	0.60	0.62	2
Mor69	0.59	4	0.05	0.13	0.57	0.54	0.55	3
Chi99	0.49	20	0.05	0.20	0.53	0.54	0.53	4
Kap51	0.51	17	0.05	0.08	0.24	0.17	0.20	22
Mag78	0.62	2	0.03	0.27	0.59	0.37	0.47	7
Kis93	0.52	15	0.03	0.09	0.44	0.23	0.32	11
Was80	0.58	5	0.02	0.11	0.41	0.55	0.47	6

Table 2. Scores and rankings for sample targets to Horowitz's 1949 performance of mazurka 63/3.

Query	*Target*	*Tempo*			T_S			T_d			*Dynamics*			*TD*		
		R_0	R_3	R_4	R_0	R_3	R_4	R_0	R_3	R_4	R_0	R_3	R_4	R_0	R_3	R_4
Rub39	Rub52	2	3	2	3	8	8	3	2	1	6	8	8	3	3	2
Rub39	Rub66	1	1	1	1	1	1	2	3	2	4	4	2	1	2	1
Rub52	Rub39	6	9	2	27	31	12	3	3	2	28	23	6	12	9	2
Rub52	Rub66	2	2	1	3	2	1	2	2	1	2	2	1	2	2	1
Rub66	Rub39	3	4	2	2	3	1	3	6	5	15	8	6	5	5	3
Rub66	Rub52	1	2	1	3	2	2	2	2	1	1	3	2	1	2	1
Ranking Averages (by feature)	R_0	2.5			6.5			2.5			9.3			4.0		
	R_3	3.5			7.8			3.0			8.0			3.8		
	R_4	1.5			4.2			2.0			4.2			1.7		
Overall Averages:		$R_0 = 4.97$			$R_3 = 5.32$			$R_4 = 2.70$								

Table 3. Rankings for 17/4 Rubinstein performances. Shaded numbers indicate perfect performance of a similarity metric.

bitrary but splitting the performance database into two equal halves seems the most flexible rule to use, and is used for the evaluation section later in this paper. But the cut-off point could be a different percentage, such as the bottom 75% of ranked scores, or an absolute cut-off number. In any case, it is preferable that the noise floor does not appear to have any favored matches, and should consist of uniform small blotches at all timescales in the scapeplot representing many different performers as is the example shown in Figure 6 (top left part of the figure). While Rachmaninoff and Zak have equivalent S_2 scores, Rachmaninoff's performance is able to cover 74% of the noise-floor, while Zak's is only able to cover 64%.

2.5 Type-4 Score

Type-3 scores require one additional refinement in order to be useful since performances are not necessarily evenly distributed in the feature space. The plots used to calculate the S_3 scores are still nearest-neighbor rank plots, so the absolute numeric distances between performances are not directly displayed. Unlike correlation values between two performances, S_3 scores are not symmetric: the score from *A* to *B* is not the same value as from *B* to *A*. It is possible for an outlier performance to match well to another performance closer to the average performance just because it happens to be facing towards the outlier, with the similarity just being a random coincidence.

Therefore, the geometric mean is used to mix the S_3 score with the reverse-query score (S_{3r}) as shown in Equation 4.

$$S_3 = A \Rightarrow B \text{ measurement} \quad (2)$$

$$S_{3r} = A \Leftarrow B \text{ measurement} \quad (3)$$

$$S_4 = \sqrt{S_3\, S_{3r}} \quad (4)$$

The arithmetic mean could also be used, but the geometric mean is useful since it penalizes the final score if the type-3 and its reverse scores are not close to each other. For example, the arithmetic mean between 0.75 and 0.25 is 0.50, while the geometric mean is lower at 0.43. Greatly differing S_3 and S_{3r} scores invariably indicate a poor match between two performances, with one of them acting as an outlier to a more central group of performances.

Table 2 shows several of the better matches to Horowitz's performance in mazurka 63/3, along with the various types of scores that they generate. S_0 is the top-level correlation between the dynamics curves, and R_0 is the corresponding similarity rankings generated by sorting S_0 values. Likewise, S_4 and R_4 indicate the final proposed similarity metric, and the resulting rankings generated by sorting these scores.

3 EVALUATION

When evaluating similarity measurement effectiveness, a useful technique with a clear ground-truth is to identify recordings by the same performer mingled among a larger collection of recordings.[5] Presumably pianists will tend to play more like their previous performances over time rather than like other pianists. If this is true, then better similarity metrics should match two performances by the same pianist more closely to each other than to other performances by different pianists.

3.1 Rubinstein performance matching

Arthur Rubinstein is perhaps the most prominent interpreter of Chopin's compositions in the 20th century, and luckily he has recorded the entire mazurka cycle three times during his career: (1) in 1938–9, aged 51; (2) in 1952–4, aged 66, and (3) in 1966, aged 79.

Table 3 lists the results of ranking his performances to each other in mazurka 17/4 where the search database contains an additional 60 performances besides the three by Rubinstein. The first column in the table indicates which performance was used as the query (Rubinstein's 1939 performance, for example, at the top of the first row). The *target* column indicates a particular target performance which is one of the other two performances by Rubinstein in the search database. Next, five columns list three types of rankings for comparison. The five columns represent four different extracted features as illustrated in Figure 1, plus the *TD* column which represents a 50/50 admixture of the tempo and dynamics features.

For each musical feature, three columns of rankings are reported. R_0 represents the rankings from the S_0 scores; R_3 being the type-3 scoring ranks, and R_4 resulting from sorting the S_4 similarity values. In these columns, a "1" indicates that the target performance was ranked best in overall similarity to the query performance, "2" indicates that it is the second best match, and so on (see the search database sizes in Table 1). In the ranking table for mazurka 17/4 performances of Rubinstein, the shaded cells indicate perfect performance matches by a particular similarity metric where the top two matches are both Rubinstein. Note that there is one perfect pair of matches in all of the R_0 columns which is found in the full-tempo feature

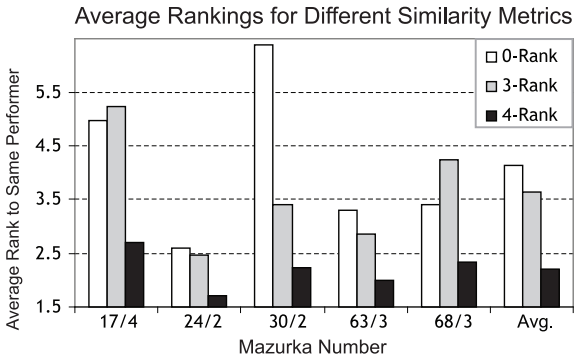

Figure 7. Ability of metrics to identify the same performer in a larger set of performances, using 3 performances of Rubinstein for each mazurka. (Lower rankings indicate better results.)

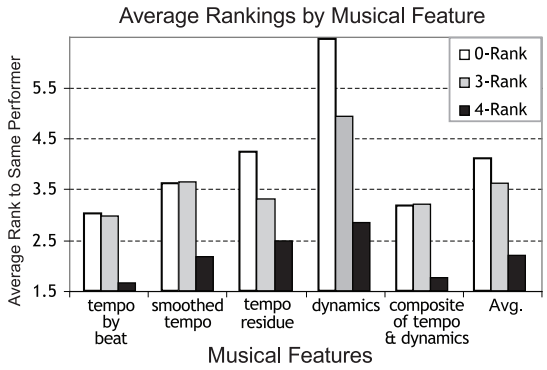

Figure 8. Ranking effectiveness by extracted musical features, using three performances of Rubinstein for each mazurka. (Lower values indicate better results.)

when Rubinstein 1939 is the target performance. No columns for R_3 contain perfect matching pairs, but about 1/2 of the R_4 columns contain perfect matches: all of the full-tempo R_4 rankings are perfect, and a majority of the desmoothed tempo and joint tempo/dynamics rankings are perfect. None of the metrics contain perfect matching pairs for the dynamics features. This is perhaps due to either (1) the dynamics data containing measurement noise (due to the difficultly of extracting dynamics data from audio data), or (2) Rubinstein varying his dynamics more over time than his tempo features, or a combination of these two possibilities.

Figure 7 shows the average rankings of Rubinstein performances for all extracted features averaged by mazurka. The figure shows S_4 scores are best at identifying the other two Rubinstein performances for all of the five mazurkas which were used in the evaluation. Typically S_4 gives three to four times better rankings than the S_0 values according to this figure. S_3 scores (used to calculate S_4 scores), are slightly better than plain correlation, but sometimes perform worse than correlation in some mazurkas.

Figure 8 evaluates the average ranking effectiveness by musical feature, averaged over all five mazurkas. Again S_4 scores are always three to four times more effective than plain correlation. S_3 scores are approximately as effective as S_0 rankings for full and smoothed tempo, but perform somewhat better on residual tempo and dynamics features, probably by minimizing the effects of sudden extreme differences between compared feature sequences caused by noisy feature data.

Mazurka	*Query*	*Target*	T R_0	T R_4	T_s R_0	T_s R_4	T_d R_0	T_d R_4	D R_0	D R_4	TD R_0	TD R_4
17/4	Cze49	Cze49b	1	1	1	1	1	1	3	1	1	1
17/4	Cze49b	Cze49	1	1	1	1	1	1	7	1	1	1
63/3	Fri23	Fri30	1	1	1	1	1	1	1	1	1	1
63/3	Fri30	Fri23	1	1	1	1	1	1	1	1	1	1
17/4	Hor71	Hor85	2	1	13	1	2	1	1	1	2	1
17/4	Hor85	Hor71	1	1	1	1	1	1	1	1	1	1
30/2	Fou78	Fou05	2	1	1	1	13	17	2	2	2	1
30/2	Fou05	Fou78	1	1	1	1	2	6	3	2	2	1
63/3	Uni32	Uni71	1	1	1	1	5	1	1	1	1	1
63/3	Uni71	Uni32	1	1	1	1	1	1	1	1	1	1

Table 4. Performer self-matching statistics.

3.2 Other performers

Rubinstein tends to vary his performance interpretation more than most other pianists. Also, other performers may tend to emulate his performances, since he is one of the more prominent interpreters of Chopin's piano music. Thus, he is a difficult case to match and is a good challenge for similarity metric evaluations.

This section summarizes the effectiveness of the S_0 and S_4 similarity metrics in identifying other pianists found in the five selected mazurkas for which two recordings by the same pianist are represented in the data collection (only Rubinstein is represented by three performances for all mazurkas).

Table 4 presents ranking evaluations for performance pairs in a similar layout to those of Rubinstein found in Table 3. In all except two cases (for Fou) the S_4 metrics perform perfectly in identifying the other performance by the same pianist. Top-level correlation was able to generate correct matches in 75% of the cases. An interesting difference between the two metrics occurs when Hor71 is the query performance. In this case S_0 yields a rank of 13 (with 12 other performance matching better than his 1985 performance), while S_4 identifies his 1985 performance as the closest match.

Fou's performance pair for mazurka 30/2 is also an interesting case. For his performances, the phrasing portion of the full- and smoothed-tempo features match well to each other, but the tempo residue does not. This is due to a significant change in his metric interpretation: the earlier performance has a strong mazurka metric pattern which consists of a short first beat, followed by a lengthened second or third beat in each measure. His 2005 performance greatly reduces this effect, and beat durations are more uniform throughout the measure in comparison to his 1978 performance.

Finally, it is interesting to note the close similarity between Uninsky's pair of performances listed in Table 4. These two performances were recorded almost 40 years apart, one in Russia and the other in Texas. Also, the first was recorded onto 78 RPM monophonic records, while the later was recorded onto 33-1/3 RPM stereo records. Nonetheless, his two performances indicate a continuity of performance interpretation over a long career.

feature	*Cortot ConA*		*Cortot Sony*	
	R_4	S_4	R_4	S_4
T	1. Rangell 2001	0.65	1. Luisada 1990	0.40
	2. Uninsky 1971	0.47	2. Ferenczy 1956	0.31
	3. Yaroshinsky 2005	0.44	3. Rubinstein 1966	0.27
	...		...	
	31. Cortot Sony	0.04	33. Cortot ConA	0.04
T_s	1. Pobłocka 1999	0.60	1. Lushtak 2004	0.27
	2. Luisada 1990	0.60	2. Milkina 1970	0.18
	3. Mohovich 1999	0.56	3. Fou 2005	0.18
	...		...	
	26. Cortot Sony	0.07	24. Cortot ConA	0.07
T_d	1. Rangell 2001	0.79	1. Luisada 1990	0.53
	2. Uninsky 1971	0.72	2. Rubinstein 1966	0.49
	3. Yaroshinsky 2005	0.57	3. Rubinstein 1939	0.47
	...		...	
	32. Cortot Sony	0.03	35. Cortot ConA	0.03
D	1. Brailowsky 1960	0.38	1. Fliere 1977	0.46
	2. Biret 1990	0.25	2. Sofronitsky 1960	0.42
	3. Milkina 1970	0.24	3. Ashkenazy 1981	0.40
	...		...	
	33. Cortot Sony	0.06	32. Cortot ConA	0.06
TD	1. Rangell 2001	0.40	1. Avg. performance	0.16
	2. Pobłocka 1999	0.37	2. Indjic 1988	0.16
	3. Yaroshinsky 2005	0.35	3. Rubinstein 1952	0.15
	...		...	
	35. Cortot Sony	0.03	34. Cortot ConA	0.03

Table 5. Comparison of Cortot performances of mazurka 30/2 (m. 1–48).

4 APPLICATION

As an example application of the derived similarity metrics, two performances of mazurka 30/2 performed by Alfred Cortot are examined in this section. One complete set of his mazurka performances can be found on commercially release recordings from a 1980's-era issue on cassette tape "recorded at diverse locations and dates presumably in the period of 1950–1952."[1] These recordings happen to be issued by the same record label as the recordings of Joyce Hatto, which casts suspicion on other recordings produced on that label.[4]. A peculiar problem is that no other commercial recordings exist of Cortot playing any mazurka, let alone the entire mazurka cycle.

In 2005, however, Sony Classical (S3K89698) released a 3-CD set of recordings by Cortot played during master classes he conducted during the late 1950's, and in this set, there are six partial performances of mazurkas by Cortot where he demonstrates how to play mazurkas to students during the class. His recording of mazurka 30/2 on these CDs is the largest continuous fragment, including 75% of the entire composition, stopping two phrases before the end of the composition.

Table 5 lists the S_4 scores and rankings for these two recordings of mazurka 30/2, with the Concert Artist's rankings on the left, and the Sony Classical rankings to the right. The five different musical features listed by column in previous tables for Rubinstein and other pianists are listed here by row. For each recording/feature combination, the top three matches are listed, along with the ranking for the complimentary Cortot recording.

Note that in all cases, the two Cortot recordings match very poorly to each other. In two cases, the worst possible ranking of 35 is achieved (since 36 performances are being compared in total). Perhaps Cortot greatly changes his performances style in the span of 6 years between these two recordings late in his life, although data from Tables 3 and 4 would not support this view since no other pianists has significantly alter all musical features at once, and only Fou significantly changes one musical feature between performances.

Therefore, it is likely that this particular mazurka recording on the Concert Artist label was not actually performed by Cortot. Results from further investigation of the other five partial mazurka performances on the Sony Classical recordings would help to confirm or refute this hypothesis, but the other examples are more fragmentary, making it difficult to extract reasonable amounts of recital-grade performance material. In addition, no performer in the top matches for the Concert Artist Cortot performance match well enough to have likely recorded this performance, so it is unlikely that any of the other 30 or so performers being compared to this spurious Cortot performance is the actual source for this particular mazurka recording.

5 FUTURE WORK

Different methods of combining S_3 and $S_{3\mathrm{r}}$ scores, such as measuring the intersection between plot areas rather than measuring the geometric mean to calculate S_4 should be examined. The concept of a noise floor when comparing multiple performances is useful for identifying features which are common or rare, and allows similarity measurements to be more consistent across different compositions, which may aid in the identification of pianists across *different* musical works.[6]

Further analysis of the layout of the noise-floor as seen in Figure 6 might be useful in differentiating between directly and indirectly generated similarities between performances. For example in this figure, Rachmaninoff's performance shows more consistent similarity towards smaller-scale features, which may indicate a direct influence on Horowitz's performance style. Zak's noise-floor boundary in Figure 6 may demonstrate an indirect similarity, such as a general school of performance.

Since the similarity measurement described in this paper works well for matching the same performer in different recordings, and examination of student/teacher similarities may be done. The analysis techniques described here should be applicable to other types of features, and may be useful with other underlying similarity metrics besides correlation. For example, It would be interesting to extract musical features first from the data with other techniques such as Principle Component Analysis[2] and use this derived feature data for characterizing the similarities between performances in place of correlation.

6 REFERENCES

[1] Chopin: The Mazurkas. Alfred Cortot, piano. Concert Artist compact disc CACD 91802 (2005).

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

[3] Sapp, C. "Comparative analysis of multiple musical performances," *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007. pp. 497–500.

[4] Singer, M. "Fantasia for piano: Joyce Hatto's incredible career," *New Yorker*, 17 Sept. 2007. pp. 66–81.

[5] Stamatatos, E. and G. Widmer. "Automatic identification of music performers with learning ensembles," *Artificial Intelligence*, 165/1 (2005). pp. 37–56.

[6] Widmer, G., S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic. "In search of the Horowitz factor," *AI Magazine*, 24/3 (2003). pp. 111–130.