Journal of New Music Research
2012, Vol. 41, No. 2, pp. 215–242

Routledge
Taylor & Francis Group

A Phylogenetic Approach to Music Performance Analysis

Elad Liebman[1], Eitan Ornoy[2] and Benny Chor[1]

[1]Tel-Aviv University, Israel; [2]Zefat Academic College, Israel

Abstract

This paper presents a novel algorithmic approach to music performance analysis. Previous attempts to use algorithmic tools in this field focused typically on tempo and dynamics alone. We base our analysis on ten different performance categories (such as bowing, vibrato and durations). We adapt phylogenetic analysis tools to resolve the inherent inconsistencies between these categories, and describe the relationships between performances. Taking samples from 29 different performances of two pieces from Bach's sonatas for solo violin, we construct a 'phylogenetic' tree, representing the relationship between those performances. The tree supports several interesting relations previously conjectured by the musicology community, such as the importance of date of birth and recording period in determining interpretative style. Our work also highlights some unexpected inter-connections between performers, and challenges previous assumptions regarding the significance of educational background and affiliation to the historically informed performance (HIP) style.

1. Introduction

The 120 years of recorded musical data provide a broad platform for studying interpretation profiles and their mutual influences. And yet, the analysis of sound recordings is a relatively new area in musicology. In its early stages, the examination of performance styles was mainly done manually, through meticulous aural scrutiny. However, a few precursors to the application of technology in the field of music analysis (and ethnomusicology in particular) could be traced to the seminal works of Carl E. Seashore and Milton Franklin Metfessel, who used phonophotography as well as the tonoscope (Williams, 1931; Seashore, 1938). Furthermore, as early as the 1950s, several researchers (most prominently Charles Seeger) used the melograph—a pitch analysis tool, which underwent a process of evolution over the years—for similar purposes (Seeger, 1951; Dahlback, 1958; Cohen & Katz, 1968; Cohen, 1969; List, 1974; Moore, 1974). Nowadays, the practice of recording analysis is aided occasionally by various software tools for determining specific categories, such as beat extraction and spectral analysis.[1]

Recording analysis has thus far led to a number of interesting findings, of which the major one is the identification of evident differences in performance styles manifested over the years. Analysing recordings made by any one performer over a large span of time or examining recordings made of the same repertoire throughout the 120 years of recordings has shown a huge shift in approach regarding tone production, tempo, articulation and the like (Philip, 1992, 2004; Leech-Wilkinson, 2009a). Another finding is the clear distinction observed by musicologists between the 'mainstream' performance approach, and the relatively new 'historically informed

[1]The amount of performance practice studies is wide, and only a few will be presented here. Such, for example, are studies made of interpretation approaches to piano compositions (Repp, 1992; Rink, 2001; Musgrave & Sherman, 2003), string quartet (Turner, 2004), symphonic repertoire (Bowen, 1996), singing techniques (Brown, 1997; Timmers, 2007), or violin performances (Sevier, 1981; Bomar, 1987; Field, 1999; Cseszko, 2000; Katz, 2003; Milsom, 2003; Fabian, 2005; Ornoy, 2008). An extended list of studies made on the subject can be additionally found in Bowen (2005).

Correspondence: Benny Chor, School of Computer Science, Tel-Aviv University, P.O. Box 39040, Tel Aviv 69978, Israel.
E-mail: benny@cs.tau.ac.il

http://dx.doi.org/10.1080/09298215.2012.668194

performance' (HIP) performance style. The use of period instruments as opposed to their 'modern' equivalent or the manner of execution of certain rhythmic elements are but a few examples of the substantial difference in interpretation between the two schools (Haskell, 1988; Kenyon, 1988; Taruskin, 1995; Lawson & Stowell, 1999; Butt, 2002; Rink, 2002; Fabian, 2003; Golomb, 2005; Ornoy 2006, 2007).

Furthermore, a widespread conception is that newer performances are less idiosyncratic in nature, compared to older ones. The reason for this is assumed to be the rise of the recording industry and the canonization of certain recordings made by authoritative figures, which are believed to have promoted a certain degree of unity and standardization in performance (Dart, 1954; Dreyfus, 1983; Philip, 1992, 2004; Katz, 2004). In contrast, recent studies challenging such assumptions have focused on identification of performers' individual, distinctive characteristics and idiosyncratic expression (Cook, Clarke, Leech-Wilkinson, & Rink, 2009; Fabian & Ornoy, 2009; Leech-Wilkinson, 2009a).

From an algorithmic point of view, several attempts have been made to comparatively analyse and classify performances (Beran & Mazzola, 1999; Madsen & Widmar, 2006; Sapp, 2007, 2008; Molina-Solana, Lluís Arcos, & Gomez, 2008; Almansa & Delicado, 2009). These efforts focused mainly on just two performance aspects (dynamics and tempo), and commonly utilized statistical and signal-processing approaches in order to compare performances to one another. It should be noted that the dynamics aspect alone is potentially problematic, as it is heavily dependent on recording technique and equipment, as well as manual intervention by the recording technician—more so than other performance aspects (Trapani & Richter, 1985; Nannestad, 2004; Trezise, 2009).

In this work, we propose a novel algorithmic approach to comparative musical recording analysis. We study 29 performances of two of Bach's sonatas for violin solo (specifically, the opening segments of the first movements of sonatas BWV 1001 and BWV 1005; see Figures 1 and 2). Applying phylogenetic reconstruction tools, we build an unrooted tree, whose 29 leaves correspond to the performances.

We consider eleven categories, such as vibrato, tempo, and chord types (see expanded method section for the full list of categories). Each performance is encoded as a 87-dimensional vector by sampling these 10 categories from predetermined, synchronous segments. These segments span 15 bars ($\sim$75 s on average) from two specific movements, chosen for their highly expressive and informative characteristics. The different categories are essentially incomparable and inconsistent, and therefore do not induce a single, uniform distance measure. To analyse them, we partition the vectors' coordinates by categories, and construct quartets (Strimmer & von Haeseler, 1996; Chor, 1998), based on each category, and a choice of 'quartets parameters'. A phylogenetic quartet is a topological arrangement of four items partitioning them into two disjoint pairs. Quartets are useful in cases in which the 'larger picture' may not be immediately deduced from the raw data, but on smaller scales, local relations may be discerned more reliably (see Figure 3).

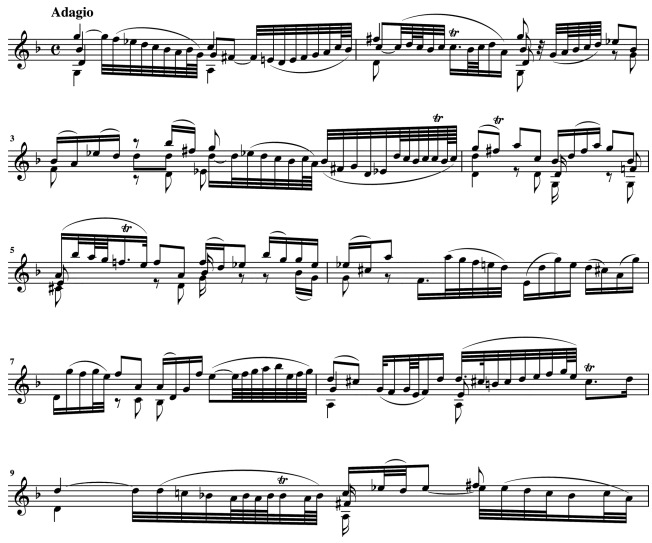

Fig. 1. Adagio (bars 1–9) from J.S. Bach's g minor Sonata (no. 1) for Solo Violin, BWV 1001. (J.S. Bach: Sechs Sonaten Und Partiten Fuer Violin solo, Urtext, based on Bach's autograph score.)

In our work, we retain only the quartets which are identified as highly reliable and combine the resulting quartet set into a tree, using a quartet max-cut heuristic (Snir & Rao, 2006). Different sets of quartets, corresponding to different choices of parameters, give rise to different trees, which are eventually combined into one final tree, using consensus (Adams, 1972). The consensus tree is then analysed to examine proximity relations between leaves, and how they relate to specific criteria, such as recording dates, performers' dates of birth, performers' music schools, and affiliation with the 'HIP' style. We note that these categories are based upon observations discussed earlier in the literature (Philip, 1992, 2004; Rink, 2002; Fabian, 2003; Katz 2004; Golomb, 2005; Ornoy, 2008; Leech-Wilkinson, 2009a).

1.1 Analysis criteria

Attempts to define performance 'style' and to explain its change over the years have been associated with a plethora of causes. Such causes include common aesthetic standards, performers' mutual influences, or shared biographical identities (Dart, 1954; Dreyfus, 1983;

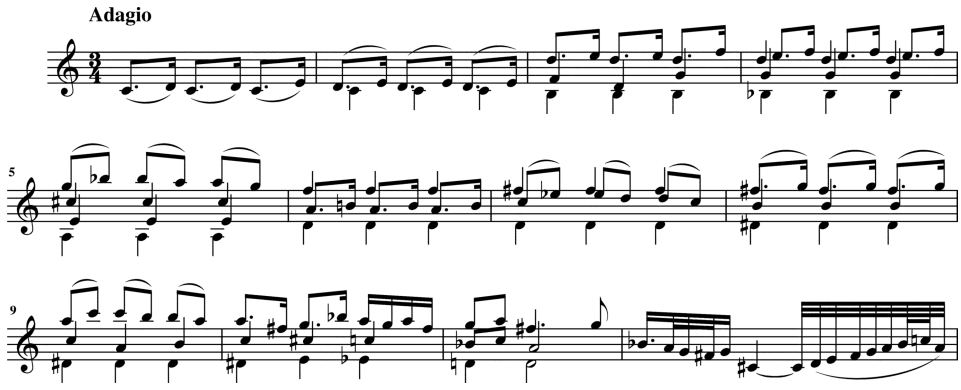

Fig. 2. Adagio (bars 1–12) from J.S. Bach's C Major Sonata (no. 3) for Solo Violin, BWV 1005. (J.S. Bach: Sechs Sonaten Und Partiten Fuer Violin solo, Urtext, based on Bach's autograph score.)

Fig. 3. A phylogenetic quartet example. For each set of four items, there are three possible quartet topologies. In this case, the four items (A,B,C,D) are partitioned by the quartet to (A,B) versus (C,D). The two other possible topologies for this set of items would be (A,C) versus (B,D) and (A,D) versus (B,C).

Philip, 1992, 2004; Weiss, 1992; Day, 2000; Katz, 2004, 2006; Lisboa, Williamon, Zicari, & Einholzer, 2005; Hellaby, 2009).[2]

The clear existence of changing yet well-defined conventional trends points to the importance of the recording date in the shaping of performance interpretation prototypes. As such, one may anticipate conformance between the date of recording and the grouping in the tree. That said, we note that performers of the second half of the twentieth century have been traditionally regarded as portraying a unified, homogenous style of performance, whereas individuality and a variety in syntax and style is viewed as dominating recordings of earlier periods. According to this concept, one would expect to find newer recordings clustered together even more so than older ones (Dart, 1954; Dreyfus, 1983; Philip, 1992, 2004; Katz, 2004).

Furthermore, since recording dates do not necessarily match birth dates, performers' age might serve as an essential factor in absorbing the influence of the recording industry on prevalent norms of practice. The invention of magnetic tape in the mid-1940s brought about an unprecedented circulation of commercial recordings. Assuming that performers' average periods of study encompass at least 20 years, it is clear that the earlier a performer was born, the more likely he is to have been educated in an era when general norms of practice might not have been influenced by recordings. On the other hand, the later a performer was born, the more likely he is to have been exposed to recordings throughout his period of study. As mentioned, homogeneity and influential prototypes of interpretations are traditionally acknowledged as predominating younger generation performers more so than their predecessors. One would therefore expect to find the youngsters clustered together, portraying more of a unified style.

Performers' schools serve as yet another biographical element, with 'school' traditionally relating to the geographic location of a music conservatory or to a particularly authoritative teacher. Nineteenth century violin school classifications include, among others, the 'French' school (Baillot, Rode, Kreutzer), 'Franco-Belgian' (Bériot, Vieuxtemps, Wieniawski, Ysaÿe) or 'German' (Spohr, David).[3] Yet such divisions seem artificial and unrelated to the existing state of affairs vis-à-vis modern violin playing, as the typical course of study for twentieth-century performers involves many different teachers throughout their years of training. As illustrated in several studies, taking into consideration the variety of master classes, courses and other related studies characteristic of a modern player's educational résumé, the significance given to one's 'school' should be considered with great caution (Boyden, 1980; Ornoy,

[2]For a comprehensive discussion of the subject see Leech-Wilkinson (2009a).

[3]See Schwarz (1977) for a review of the migration and influences of prominent Russian violinists on the 'American' school, Halász (1995) for a discussion of the Hungarian school, Lauer (1997) for the Franco-Belgian school, and Lankovsky (2009) for the Moscow violin school.

2008; Fabian & Ornoy, 2009; Leech-Wilkinson, 2009a). That is not to say, however, that some significance should not be given to direct relations, such as violinists who have studied under the same teacher or who have been analysed together with their pupils. In such cases, to a certain degree, one could expect to find congruence to the location in the tree.

The term 'historically informed performers' (HIP) is commonly used to describe the large group of musicians who perform early music repertoire in the 'authentic' manner in which it was historically written and performed. They are commonly distinguished from their 'mainstream' colleagues, i.e. performers who apply 'modern' performance practices, which seem incompatible with the goals of the HIP movement. Previous studies have discussed the homogeneity in the execution of central musical parameters found among HIP, such as similar manners of rhythmic interpretation, tempo choices or performance elements influencing sound production (Taruskin, 1995; Fabian, 2003; Golomb, 2005; Ornoy, 2006). One could thus expect that HIP performers be clustered together, indicating congruence in most parameters. It should be noted that in the context of this work, violinists belonging to the HIP category are those who have used period instruments during recording. Instances where other features of 'stylistic awareness' were presented (such as the use of a curved bow or special rhythmic execution) were excluded from such classification.[4]

The four analysis criteria described above have been discussed quite extensively in the professional literature. This has motivated us to examine them with respect to our algorithmic analysis as well. We emphasize that our analysis does not make any prior assumptions or hypotheses regarding these criteria.

2. Methods summary

In this section, we describe the input data and explain the pre-processing and analysis phases. A more in-depth and technical overview of the methods employed at each stage can be found in Appendix C.

2.1 Data

We have based our analysis on 29 different recordings of Bach's sonatas and partitas for solo violin. Segments of two specific movements (BVW 1001 Adagio and BVW 1005 Adagio) were selected for data analysis. For each performance, measurements belonging to 10 distinct categories were examined. In each category, a number of features were extracted.

The 10 categories are:

- *Bowing*—the marking of bow changes (determined by auditory means). Each measurement represents whether bow direction was changed, partially changed (by the use of Portato/Louré, i.e. slight audible separation of slurred notes without changing the bow direction) or unchanged at ten diachronic points in the sampled section of BVW 1001 (Gm adagio). These points were chosen after meticulously studying the recordings, so that at each such point, at least one performer had indeed changed bow direction.
- *Chord ratio*—the ratio between the lowest and the highest notes in the sampled chords (measured with the Sonic Visualiser software package, see Section 2.2 for a detailed explanation of this category).
- *Double stop/arpeggio*—represents whether the chord is an arpeggio or a double stop (one measurement was taken for each chord in the analysis range, determined by auditory means).
- *Count of double stops in C adagio*—double stop frequency.
- *Vibrato*—split into three 'sub-features': depth, speed and onset (measured with the Sonic Visualiser software package).
- *Duration per bar (mid-phrase durations)*—measured with the Sonic Visualiser software package. These measurements were taken from the sampled segment of BVW 1005 (C adagio), since each bar in this segment is meaningful in terms of phrasing.
- *Tempo changes*—the difference between adjacent duration measurements, which were collected for the previous category. This measure is converted to a discrete scale of three values—*faster*, *slower* and *unchanged*.
- *Total duration*—overall performance duration (measured with the Sonic Visualiser software package).
- *Dotting ratio*—ratio between adjacent long and short notes (measured with the Sonic Visualiser software package, based on the first bar of the CM Adagio movement).
- *Standard deviation of the tempo changes*—the standard deviation of the tempo changes measurements. This feature is useful in quantifying the tempo variance of a given performer.

As discussed in Section 1, we preferred not to incorporate dynamics-related data into our analysis, as it is too heavily dependent on recording equipment and technique, as well as on manual intervention by the recording studio technician.

[4]Addressing the wide issue concerning 'historically informed performances' is beyond the scope of this paper. However, it should be mentioned that this categorization was based on previous studies which have pointed to the eminence of period instruments among performers connected with the HIP agenda (see Boyden, 1980; Haskell, 1988; Kenyon, 1988; Fabian, 2000, 2003; Ornoy, 2006; Haynes, 2007).

2.2 Musicological considerations

We faced a practical challenge in analysing the frequent triple and quadruple-stops contained in the analysed movements, as there are many possible ways to execute them. The different ways of breaking the chord—which are directly linked to idiomatic preferences, technical limitations, or to the chord's function in the overall musical context—often involve rhythmic alterations of its inner-notes. Two categories serve for comparison of interpretation: '*double stop/arpeggio*', which distinguishes between arpeggios and chord-breaking, and '*chord ratio*', which examines whether the highest note of the chord is more dominant than the lowest note (the duration of both notes was measured). These categories provide meaningful data regarding both the melodic and the polyphonic aspects of the interpretation (Boyden, 1965; Efrati, 1979; Lester, 1999; Katz, 2003).

Since bow change serves as an idiomatic parameter, which at present is not amenable to computerized software analysis, it was obtained through meticulous repeated aural scrutiny of the relevant recordings. Measurements were made by both the first and second author (the latter being a professional violinist) during several listening sessions.

2.3 Computational considerations

2.3.1 Processing—quartets and phylogenetic trees

The combined data was first normalized and then examined. Initially, a unified distance measure was calculated, based on the 89-long vectors representing all categories. By applying the Buneman (1971) tree criteria on the unified distance measure, we discovered that it has a highly incongruent nature—the resulting tree was completely unresolved (a star). More evidence for the implausibility of naïvely combining the categories into a single unified measure was obtained using clustering. To analyse the data, it was clustered repeatedly, each time according to a different category, using the k-means algorithm (MacQueen, 1967). Since the output of the k-means algorithm is dependent both on the parameter k and on a random initialization stage, the clusters were determined several times, using different k values ($k = 2$, 3, 4). The different partitions induced by the various categories were then compared and found to be highly inconsistent. For these reasons, classic distance based approaches (such as neighbour joining and standard clustering) were deemed inapplicable. We thus decided to adopt a quartet based approach (Strimmer & von Haesler, 1996; Chor, 1998; Ben-Dor, Chor, Graur, Ophir, & Pelleg, 1998; Jiang, Kearney, & Li, 2000).

Our quartet-based scheme is as follows: initially, we work with each category separately, choosing those quartet topologies, which have clear support. To determine support, a voting scheme is used—for each possible 4-tuple (a subset of four out of the 29 performances), each category votes for a specific topology mapping the relations between these four performances, or abstains if it supports no such topology. After this stage, quartet topologies with insufficient support (defined by number of supporting categories as well as the number of opposing categories) are filtered out, leaving us with a relatively small list of quartet topologies which are deemed reliable. This list is dependent on several predetermined parameters (defining category support and reliability). Therefore, the process is repeated many times for different parameter configurations. For each such list of quartet topologies (determined by a specific set of parameters), an unrooted phylogenetic tree is constructed, using Snir and Rao's (2006) 'quartets max-cut' heuristic. The problem of building a tree from quartets is computationally intractable, thus a heuristic is called for (Steel, 1992).

Subsequently, each tree is given a score, based on its rate of accordance with the list of quartet topologies from which it was constructed, as well as the size of this 'support list' and the number of splits the resultant tree contains (we give preference to trees which are based on a large number of quartet topologies, and trees which are resolved enough to display meaningful information).

2.3.2 Processing—consensus trees, final tree selection

Having scored all the trees created by the various quartet topology lists (essentially covering the parameter space), a list of majority-vote consensus trees (Margush and McMorris, 1981) is constructed—for the 20, 40, 60, 80 and 100 highest scoring trees (out of the list of meaningful trees we described earlier).

The five resultant consensus trees are quite similar, as can be seen in the Table 1, presenting the pairwise distances between the trees according to the Robinson–Foulds metric (Robinson & Foulds, 1980). In this sense, we can say that the resultant consensus trees are stable.

In order to compare the various consensus trees, we devised two 'tree quality' measures, used to determine which tree is most reliable. The selected tree was cons_80Trees (the tree constructed via consensus over the 80 highest scoring trees). The second best tree was cons_20Trees (the tree constructed via consensus over the 20 highest scoring trees). We note that cons_80Trees obtained better results than cons_20Trees for approximately 60% of the quartet lists considering the first quality measure, and approximately 80% considering the second quality measure, thus indicating the resultant tree is consistently superior to the 'competing' option.

Table 1. Robinson–Foulds distances of the five resulting trees (*the maximal distance score in our case is* $2 * 28 = 56$).

	Cons_20Trees	Cons_40Trees	Cons_60Trees	Cons_80Trees	Cons_100Trees
Cons_20Trees	0	–	–	–	–
Cons_40Trees	8	0	–	–	–
Cons_60Trees	8	0	0	–	–
Cons_80Trees	14	6	6	0	–
Cons_100Trees	10	2	2	4	0

2.4 Correlating the distance and the categories

We attempted to discern which performance categories were most influential in the tree construction process. Each category is represented by a vector whose length varies between 1 (for the total duration category, for instance) and 30 (for the double stop versus arpeggio category). For each pair of performances, the Euclidean distance between the vectors of each category was computed. For each of our 10 categories, we proceeded to compute the Pearson correlation between the Euclidean pairwise distances among performances induced by this category, and the tree distances between these performances. Overall, this is the Pearson correlation between two vectors of length $\binom{29}{2} = 406$ each. We also computed this correlation for the distance induced by all the categories naïvely concatenated. There are 10 categories, plus the 'combined' one, so overall, 11 correlations were computed. All correlation results were strictly positive. Three of the eleven resulting values were in the range $[0.04, 0.1)$. Four of these values were in the range $[0.1, 0.2)$. Two correlation values were in the range $[0.2, 0.3)$, and the top two correlation values were 0.4 and 0.42. These correlations correspond to the count of double stops in the first movement, and the dotting ratio, respectively. This means that these two categories are the most correlated to the resulting tree, in terms of pairwise distances. Interestingly, the correlation of the distance induced by the 'combined' (concatenated) category is only 0.12. This may serve as further evidence for the non-metric nature of the data, explaining why naïve neighbour joining failed to produce an informative tree.

2.5 Validation

The inhomogeneous nature of the data makes validation a non-trivial task, as the validation criteria themselves are not clear a priori. Still, we would expect a considerable correlation between the raw input data and the resultant tree. For this purpose, we used the entire list of supported quartets (i.e. all the quartets generated by each of the ten categories, with respect to at least one of the thresholds), and devised a simple measure of pair support/opposition—how strongly is the entire 'mass' of supported quartets (for all categories) 'in favour of' placing a given pair of performances together versus putting them apart. It should be noted that a lack of strong enough support in favour of placing two performances together does not necessarily entail that the data supports placing them apart—the input data may be undecided for specific pairs. The votes were summed to constitute an opposition score and a support score (the two scores were calculated separately for the reason listed above). In addition, the actual distances in the resultant tree were calculated. Then, the Pearson correlation between the tree distances (for each pair of performances, $\binom{29}{2} = 406$) and the rates of support/opposition described above was calculated (two correlation scores, separately). The correlation between the pairwise performance tree distances and the opposition score is $+0.557$, whereas the correlation between the pairwise performance distances and the support score is -0.556. The proximity between the positive and negative correlation values should not come as a surprise, as the correlation between support and opposition was calculated to be -0.92 (and not –1, due to the 'indifferent' categories).

This illustrates that the tree distances are well correlated to the input. On the other hand, it also suggests that the final result is highly sensitive to many other factors not accounted for by our measure (such as the mutual information among quartets voting for a given pair, or against it). Therefore support and opposition do not fully determine the final positions observed in the tree.

3. Results

In this section we describe the resulting tree, and analyse its correspondence to four criteria that are commonly used in music-performance studies for recording analysis. In addition, we show that inconsistencies in the data make standard clustering analysis problematic.

Our analysis yields the unrooted phylogenetic tree depicted in Figure 4. The resultant tree has 29 leaves (the performances), contains 27 splits, and its diameter is 13 edges long. All splits in the tree are binary. Of the 29 performances, three pairs are by the same performer. Of these, the two performances by Heifetz (from 1935

Fig. 4. The resultant tree. The numbers refer to subtree numbering in the text.

and 1952) are siblings in the tree. The two performances by Kuijken (from 1983 and 1999) are fairly close to one another (both placed in a subtree of seven leaves), whereas the two performances by Milstein are relatively far apart—the smallest subtree containing both performances has 19 leaves).

In terms of hierarchical clustering, there are six discernable clusters (subtrees with relatively small distances between performers) induced by the resultant tree. These subtrees are marked and numbered in Figure 4. The first three clusters constitute a larger, 'upper' subtree, and the last three clusters constitute a 'lower' subtree accordingly.

As mentioned, we applied a number of standard criteria to examine this tree. For two of the criteria examined, the date of birth of the performer and the recording date, the level of agreement with the tree topology is good. For the criteria of 'main stream' versus 'HIP' performance style, there is only reasonable though less distinct agreement. For the performers' school criteria there is an overall disagreement with the tree, even though it does seem to play some minor part on local scales.

There is a fairly good correspondence between the date of birth and the location in the tree. In particular, there are two distinct subtrees well characterized by age. The first (bottom) subtree contains 13 performers (14 recordings), 10 of whom were born before 1930. Only two out of the 14 performances of those born before 1930 were placed outside this subtree—one performance (out of two) by Milstein, born 1903, and one by Telmanyi. The average date of birth for performers in this subtree is approximately 1918. The second (top) subtree contains 14 performers (15 recordings), 13 of which were born after 1941, and seven of which born after 1952. Only two performers born after 1952, Szenthelyi (born 1952), and Ehnes (born 1976), were placed outside this subtree. The average year of birth for this subtree is 1945.

We use conditional probabilities to quantify more accurately the extent of accordance between the date of birth and the placement in the tree. The probability to be born after 1941 given that the performance had been placed in the upper ('young') subtree is 0.866, whereas the conditional probability to be born prior to 1941 under the same assumption is 0.133. Similarly, the probability to be placed in the 'young' subtree given that the performer was born after 1941 is 0.812, whereas the probability to be placed in the opposite subtree given the same assumption is only 0.188 (see Table 2).

A more refined resolution is difficult to obtain, as the tree does not globally betray the exact location of a specific performance according to the performer's date of birth. This is understandable as age alone cannot be expected to determine a performer's expressive signature, and taking into account the huge variance of birth and recording dates in the tree.

However, on average, there is a strong correlation between tree distances and differences in dates of birth. We have partitioned the $\binom{29}{2} = 406$ age differences to nine bins: those pairs whose performers' dates of birth are at most 5 years apart, those whose dates of birth are between 5 and 15 years apart, and so forth, up to the last bin, which represents pairs of performances whose performers' dates of birth are 75 years apart or more. For each such bin, we calculate the corresponding distances in the tree, and calculate the average tree distances within this bin. Then, we computed the Pearson correlation between the average tree distance within the bins and the average dates of birth difference within the bins and found it to be 0.697. So indeed, the more distant two performers are in terms of eras of activity, the more distant they are likely to be in the resultant tree. Additional evidence to this relation is obtained if we examine it from the opposite direction: examining the tree distances, partitioned similarly into four bins (whose centres are roughly 3, 6, 9, and 12). The correlation between the average tree distance within these bins and the average date of birth difference is 0.947. Figures 5 and 6 display these relations graphically.

The analysis for recording dates yields fairly similar results. Out of the 29 recordings examined in this work,

Table 2. List of performances, sorted by performer's date of birth, with affiliation to 'elder'/'junior' major subtree and specific subtree. Partition is by year of birth (before/after 1930).

Performer's name	Date of birth	Recording date	Found in 'Elder' (bottom) vs. 'Junior' (top) subtree	Association to specific subtree
Enescu	1881	1948	Elder	6
Szigeti	1892	1931	Elder	4
Telmányi	1892	1954	Junior	3
Heifetz	1901	1935	Elder	5
Heifetz	1901	1952	Elder	5
Milstein	1903	1954	Elder	4
Milstein	1903	1975	Junior	2
Végh	1905	1971	Elder	4
Menuhin	1916	1957	Elder	6
Szeryng	1918	1968	Elder	6
Ricci	1918	1981	Elder	6
Grumiaux	1921	1960–1961	Elder	6
Suk	1929	1971	Elder	6
Gahler	1941	1998	Junior	2
Luca	1943	1977	Junior	3
Kuijken	1944	1983	Junior	1
Kuijken	1944	2001	Junior	1
Perlman	1945	1986	Elder	4
van Dael	1946	1996	Junior	1
Kremer	1947	2005	Junior	1
Szenthelyi	1952	2002	Elder	6
Wallfisch	1952	1997	Junior	1
Hugget	1953	1995	Junior	3
Mintz	1957	1984	Junior	2
B.Brooks	1959	2003	Junior	2
Zehetmair	1961	1983	Junior	1
Tetzlaff	1966	1994	Junior	1
Podger	1968	1999	Junior	3
Ehnes	1976	1999–2000	Elder	6

11 are dated prior to 1971. 10 out of these 11 recordings are placed on the lower (or 'senior') subtree discussed earlier (containing 14 recordings all in all). The average recording date for the 'senior' subtree is approximately 1965. Out of the 18 recordings done after 1971, 14 were placed in the 'young' subtree (consisting of 15 recordings overall). Fourteen out of 15 performances in the 'young' subtree were recorded after 1975, and the average recording date for this subtree is approximately 1990. The conditional probability of being placed in the 'young' subtree given that the recording was made after 1971 is 0.778, whereas the probability of being placed in the 'senior' subtree given the same assumption is 0.222. Similarly, the probability of being recorded after 1971 given that the recording was placed in the 'young' subtree is 0.933, whereas the probability to be recorded prior to 1971 given the same assumption is only 0.067 (see Table 3).

We note that, as expected, the date of birth and the recording date are highly correlated (a Pearson correlation of +0.87 was found), which easily explains why the tree behaves similarly under both measures.

As explained regarding the date of birth criterion, achieving a finer predictive resolution is unlikely, but again, there is a strong correlation between the difference in recording dates and the average tree distance. If we partition the pairwise recording date differences to bins whose centres are roughly 0, 10, ..., 70, the Pearson correlation between the average recording date difference and the average tree distance is 0.73. In the other direction, the correlation between the average tree distances (partitioned to bins whose centres are roughly 3, 6, 9, and 12) and the average recording date difference is 0.643 (notice this correlation is weaker than that observed for dates of birth). Figures 7 and 8 display these relations graphically.

Figure 9 depicts the resulting performance tree with an emphasis on the HIP performances (underlined). There are eight such performances, all of which are placed with relative proximity across the upper subtree (eight out of the 15 performances in the upper, 'young' subtree are considered HIP). All in all the agreement of the tree with this category is only moderate.

Figure 10 depicts the performers' alleged affiliation to musical schools (roughly divided into five categories) based on their primary teachers (see Appendix B and Ornoy (2008) for a comprehensive overview of performers' schools). Note that many performers are associated with two schools. These labels are not localized in the tree structure—namely, there is no discernable agreement between the location of a performance in the tree and its association to musical schools. However, it should be noted that if we examine this category on small, local scales, such as pairs of sibling performances in the tree, many of them do share a school affiliation. Out of the eight sibling pairs in the tree (in addition to the two Heifetz performances which were also paired) five have common schools (namely Szigeti and Vegh, Kuijken and Wallfisch, Zehetmair and Kremer, Brooks and Gahler, and Hugget and Telmanyi). If we consider 'cousin'–'uncle' relations as well, 4 out of 14 such relations are also supported by their school affiliation (namely Tetzlaff and Van Dael, Suk and Szerying, Enescu and Menuhin, and Grumiaux and Enesco).

Table 4 represents a summary of the results, according to the four criteria described.

3.1 Clustering analysis

An attempt was made to analyse how the performances relate to one another, based on single categories and category pairs. A clustering approach was utilized, using the k-means algorithm (MacQueen, 1967). It was conclusively revealed that the various performances do

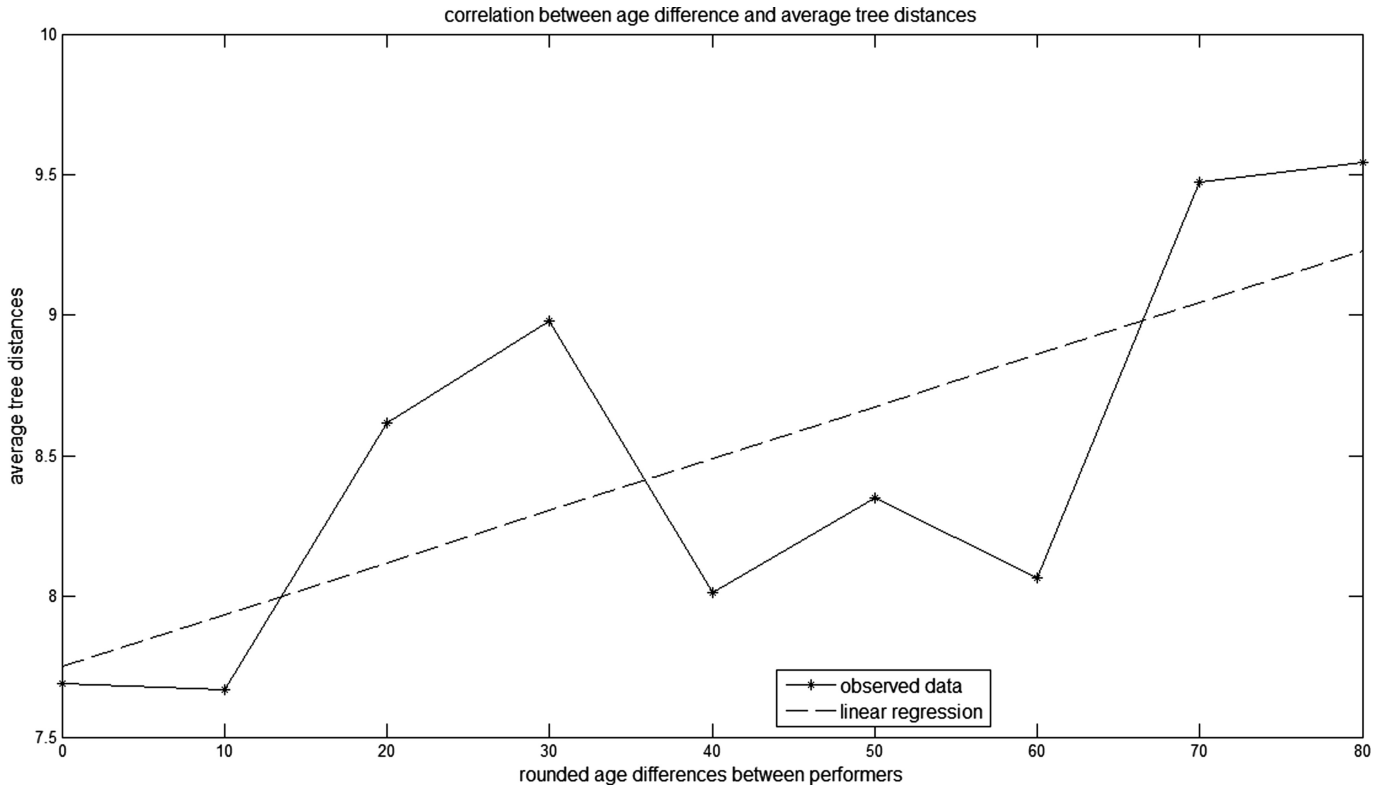

Fig. 5. Age difference versus average distance in the tree (per nine bins).

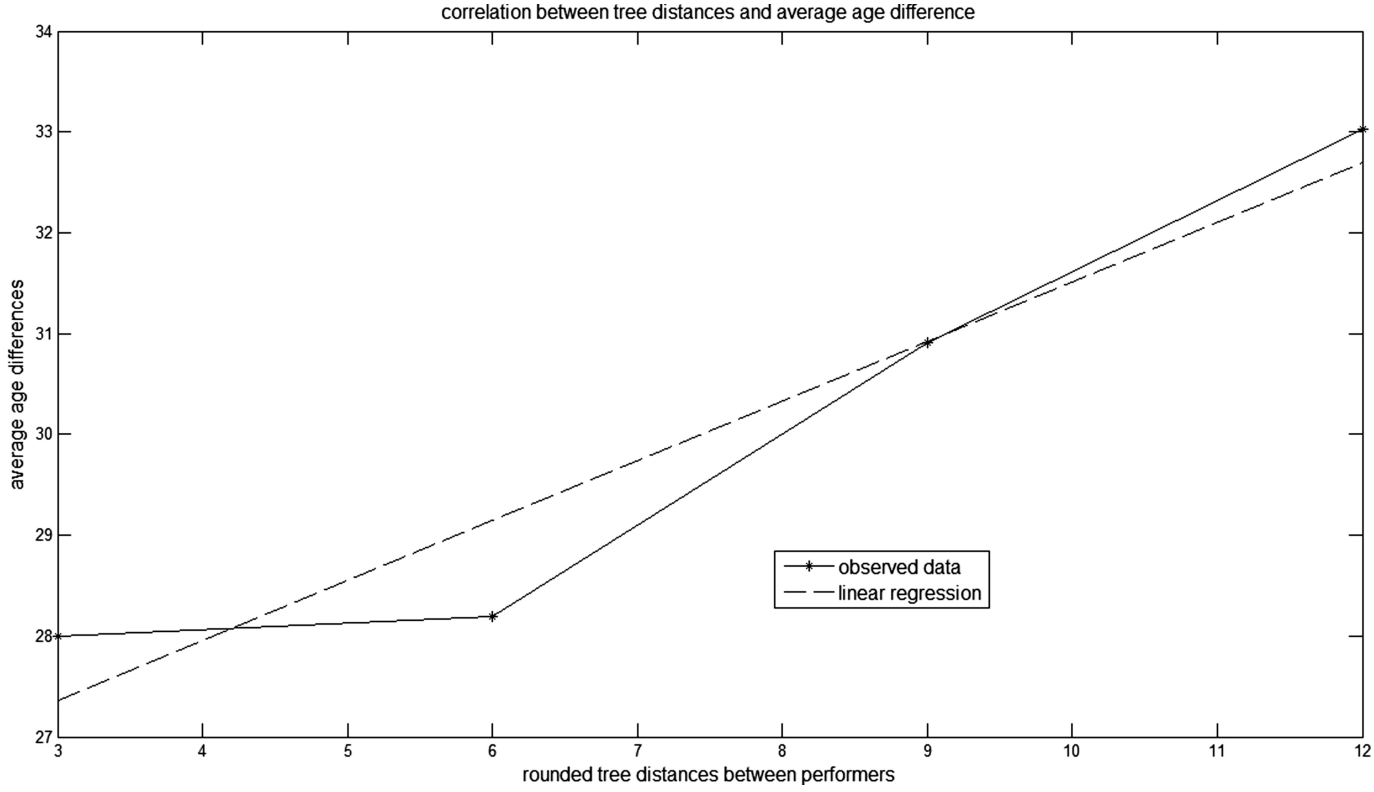

Fig. 6. Distance in the tree versus average age difference (per four bins).

Table 3. List of performances, sorted by recording date, with affiliation to 'elder'/'junior' major subtree and specific subtree. Partition is by year of recording (before/after 1971).

Performer's name	Date of birth	Recording date	Found in 'Elder' (bottom) vs. 'Junior' (top) subtree	Association to specific subtree
Szigeti	1892	1931	Elder	4
Heifetz	1901	1935	Elder	5
Enescu	1881	1948	Elder	6
Heifetz	1901	1952	Elder	5
Telmányi	1892	1954	Junior	3
Milstein	1903	1954	Elder	4
Menuhin	1916	1957	Elder	6
Grumiaux	1921	1960–1961	Elder	6
Szeryng	1918	1968	Elder	6
Végh	1905	1971	Elder	4
Suk	1929	1971	Elder	6
Milstein	1903	1975	Junior	2
Luca	1943	1977	Junior	3
Ricci	1918	1981	Elder	6
Kuijken	1944	1983	Junior	1
Zehetmair	1961	1983	Junior	1
Mintz	1957	1984	Junior	2
Perlman	1945	1986	Elder	4
Tetzlaff	1966	1994	Junior	1
Hugget	1953	1995	Junior	3
van Dael	1946	1996	Junior	1
Wallfisch	1952	1997	Junior	1
Gahler	1941	1998	Junior	2
Podger	1968	1999	Junior	3
Ehnes	1976	1999–2000	Elder	6
Kuijken	1944	2001	Junior	1
Szenthelyi	1952	2002	Elder	6
B.Brooks	1959	2003	Junior	2
Kremer	1947	2005	Junior	1

not cluster well when examined as a whole—within single categories, there is no clear partition to clusters, and clusters generated by different categories are incompatible. Indeed, it was evident that the various categories are in complete disagreement regarding the division to clusters. This observation is in agreement with the non-metric, non-homogenous nature of the performances. We created two-dimensional plots of the data, each plot according to a different pair of categories. For each such plot, each axis represents a different feature category (for multivariate categories the primary dimension produced by principal components analysis was taken). These plots allow us to examine how well clusters induced by one category hold when examined by a different category. An example of this form of visual display is given in Figures 11 and 12.

This form of visual display reveals not only how the data is clustered according to each category alone, but also the level of agreement between two specific categories with respect to their clustering partition. If the two categories generally agree on the clustering partition, two-dimensional clusters should appear. If the two categories disagree, the data should appear well partitioned according to one axis, but poorly partitioned according to the other. This is illustrated quite well in Figures 11 and 12, which depict the clustering result with respect to the dotting ratio category and the overall duration category. In Figure 11, the data is clustered according to the total duration, whereas in Figure 12, the data is clustered according to the dotting ratio. It is apparent that the two clustering divisions are in poor agreement—while the data is clearly well divided according to one category, it is completely undivided according to the other. It is also evident that the division to clusters, even in each category alone, is quite arbitrary, on a global scale. For instance, while Kuijken, Wallfisch, Zehetmair, and Van Dael are relatively close together, they are quite distant from Podger, Hugget, Luca, and Brooks, in at least one of the categories. It is worth noting that while globally the categories are incompatible, some local proximity relations do make sense (such as the relative proximity between the two Heifetz performances).

Clustering analysis of other category pairs yielded similar insights (see Appendix D for two more examples).

4. Summary and discussion

4.1 Computational aspects

There are several fundamental problems one must reckon with when initiating a work of this nature. Which musical categories (i.e. performance aspects) should be taken into account, and which should be ignored? How should the chosen categories be encoded, and how can they be combined meaningfully so that their relations are revealed? Our chosen categories are accepted as significant ones in musicological literature (Brown, 1997; Fabian, 2003; Katz, 2003, 2006; Philip, 2004; Fabian & Ornoy 2009; to name a few), even though other choices might make sense as well. A problem of a different nature is that most categories cannot at the moment be retrieved by fully automatic means.

Once each category is sampled and encoded, it is desirable to combine all categories' encodings into one vector, which induces one metric. However, it turns out that these categories are incompatible, and such unification does not seem possible. Therefore, a naïve approach (such as calculating a single distance metric based on the vector of all the encoded categories in its entirety, and utilizing some classic type of neighbour joining to construct a phylogenetic tree) is bound to produce unreliable results. This led us to work separately on different category groups. Quartets from each category were generated and

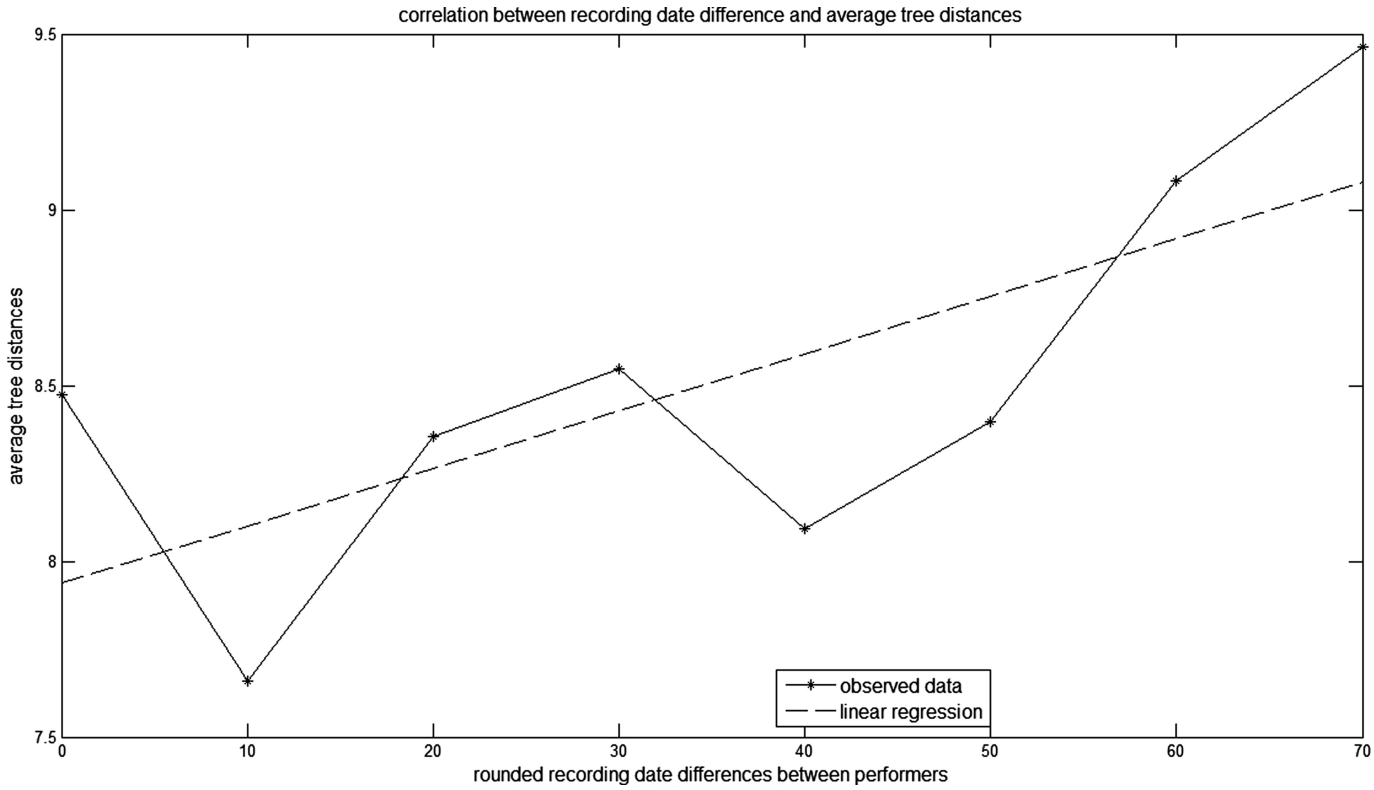

Fig. 7. Recording date difference versus average distance in the tree (per eight bins).

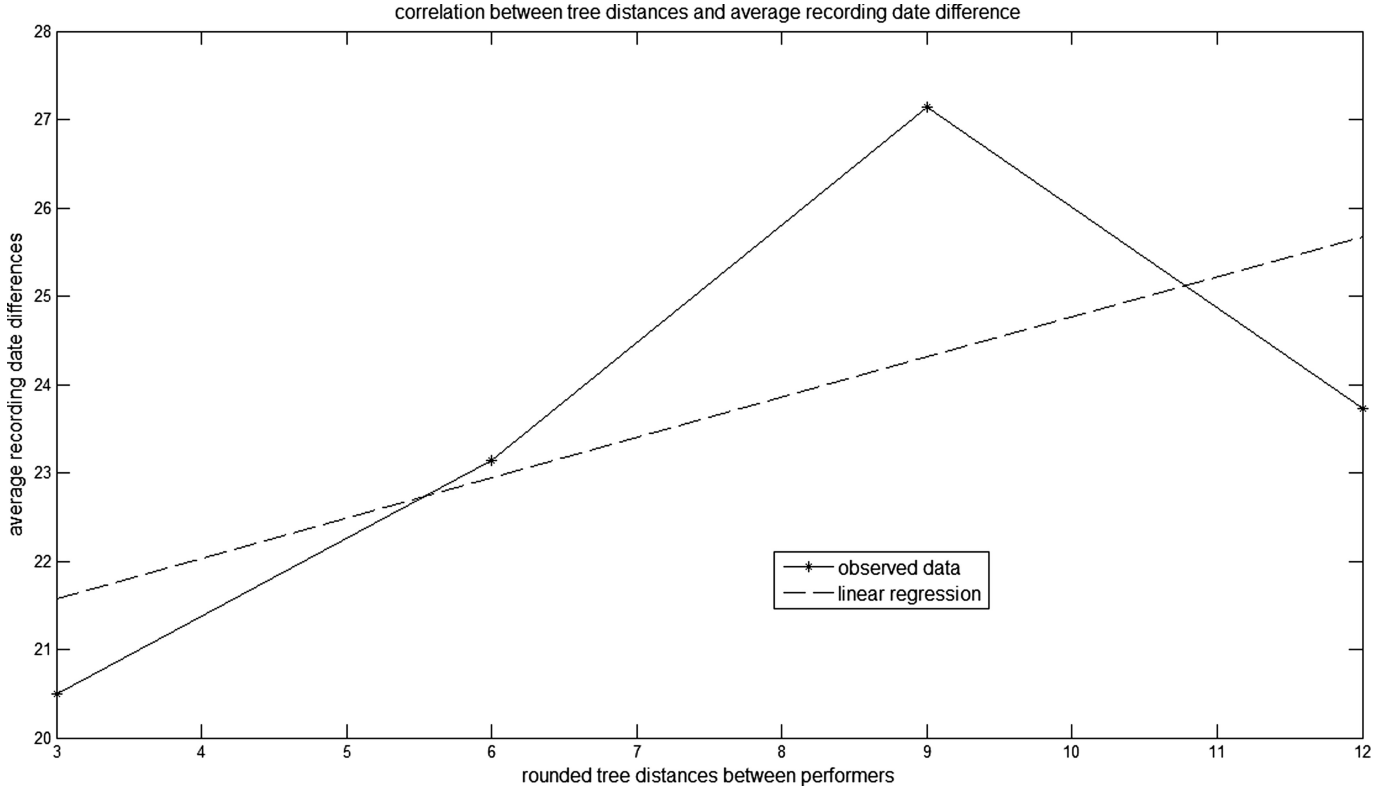

Fig. 8. Average distance in the tree versus average recording date difference (per four bins).

Fig. 9. Marking of HIP affiliated performances in the tree.

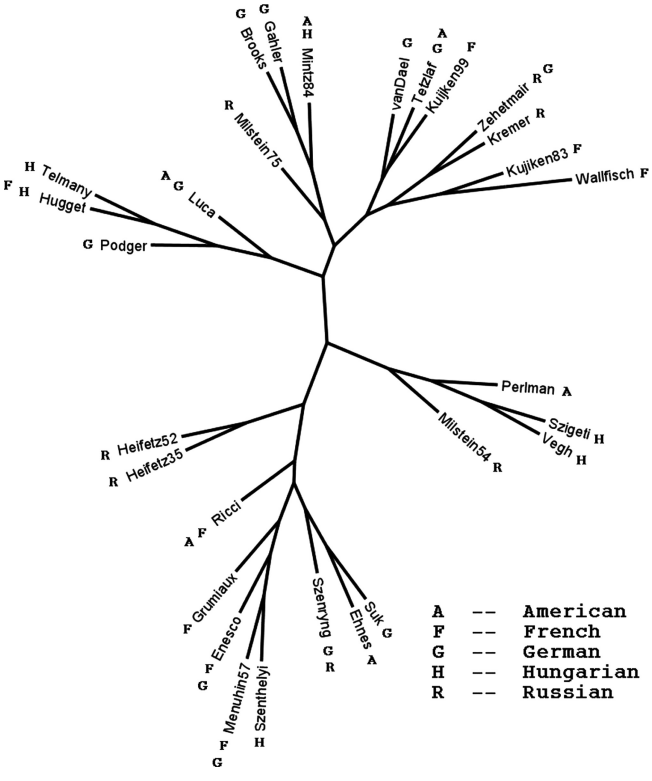

Fig. 10. Affiliation to performance schools in the tree.

Table 4. Summary of analysis by categories.

Category	Findings	Support by tree
Division between 'HIP' and mainstream performers	'Mainstream' & HIP performances are located in the same upper subtree (consisting of 15 performances)	Moderate
Division by age	Clear division to a 'senior' subtree and a 'youngster' subtree	Very good
Division by recording date	Similar division between 'old' and 'contemporary' recordings	Very good
Division by schools	No apparent clustering of performers according to school affiliation, on an overall level. Some local effect may be observed	Low

then filtered, based on the ratio between their edge lengths, retaining only quartets that are reliable.

For every set of four 'species' (performances), all quartets w.r.t. the different categories went through a 'voting' phase—each category provided a vote for a quartet topology supported by it (a category could 'abstain' if no quartet topology was convincingly satisfied by it). After this phase, the four said species are either represented by one final quartet topology or by no quartet at all (if the voting phase could not produce a convincing 'winner'). The set of surviving topologies is given as input to a 'tree from quartets' heuristic (Snir & Rao, 2006), and a tree is produced. Trees resulting from different parameters are then reconciled, using a tree consensus algorithm.

We note that in principle, one may represent the relations between performances not as a tree, but rather as a network (Huson & Bryant, 2005). Such an approach has interesting ramifications, which, we believe, should be explored further, but are out of the scope of the current work.

4.2 Musicological aspects

From a musicological perspective, this study leads to several interesting conclusions. First among our findings is the evident amalgamation of similar birth dates. As mentioned in the introduction, it is premised on the supposition that performers' age serves as an essential factor in absorbing the influence of the recording industry on conventional practices, as the overwhelming increase of commercial LPs from the mid-1940s onwards clearly acted upon newer generations of performers more so than upon their older peers. The high level of agreement with the tree topology found here clearly

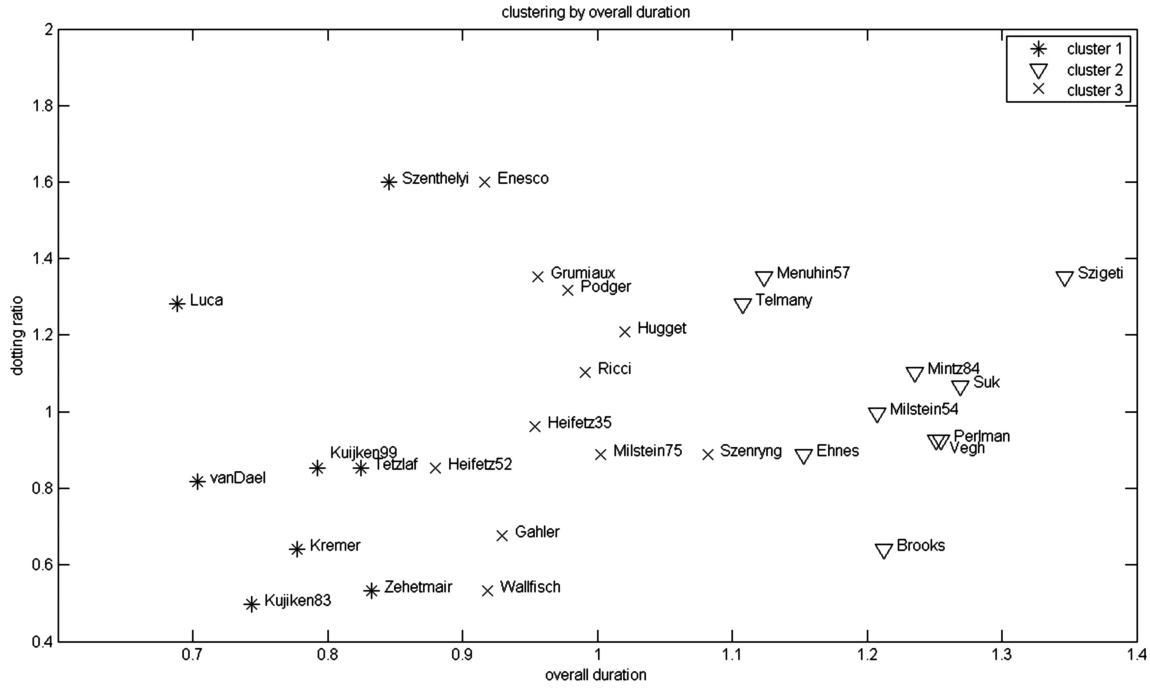

Fig. 11. Overall duration versus dotting ratio categories. Clustering according to the overall duration category (three clusters).

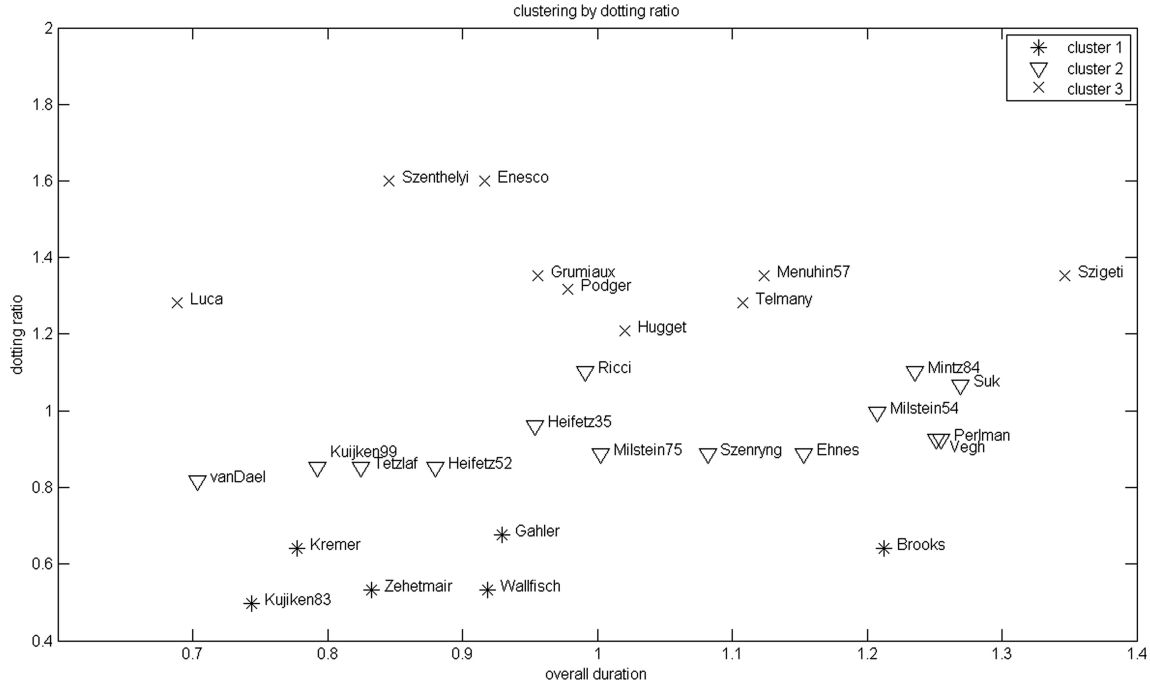

Fig. 12. Overall duration versus dotting ratio categories. Clustering according to the dotting ratio category (three clusters).

supports such a premise: findings indicate that the average date of birth for the 'older' group sub-tree is 1923, and detect a parallel sub-group of 'youngsters' in the opposite pole, whose average year of birth is 1947.[5]

Our results question the claim that older generation performers, who were less exposed to recordings by other performers during their period of training, would display more idiosyncratic characteristics in their performances.

[5]It should be noted that the 'younger' subtree is highly correlated with the 'historically-informed' subtree. This should not come as a surprise, as the first recording on period instrument of Bach's solo violin set was made by Luca in 1977.

Our tree does not support such a hypothesis, as older generation recordings are quite clearly clustered together regardless of their schools, implying general similarities in performance style rather than idiosyncrasies.

A different hypothesis asserts that recording date has a strong correlation to the performance style. As noted, recording dates do not match birth dates (but are obviously correlated with them—a Pearson correlation of +0.87 was found). When examining the date of recordings for the performances in our tree, our analysis has indeed revealed a rather strong accordance between the date of recording and the grouping in the tree. This correlation, however, is not as strong as that observed for the date of birth, which potentially may serve to illustrate the primacy of interpretive traits shaped early in one's artistic development over norms of practice emerging after his formative years. This finding corresponds to similar observations stating the importance of birth date, rather than recording date, on performers' individual style throughout the years.[6]

We note that previous analyses of different recordings made by the same artist do suggest stylistic change over the course of time.[7] This observation has partial support in our tree, as our input includes three pairs of recordings by the same performer from different years (recordings by Heifetz, Milstein, and Kuijken). While the two Heifetz recordings are placed as siblings (insinuating consistency in the manner of execution between his 1935 and 1952 recordings), the two recordings made by Milstein and Kuijken are placed farther apart from each other. And indeed, one may notice variation when comparing their categories—for instance, the vibrato and overall tempo, the dotting ratio, as well as the chord type and the tempo variance, are quite different in Milstein's two performances.[8]

Obviously, while performers' early interpretive imprint is of extreme significance, artists may still adapt to changing aesthetics and influences over the years. One should also note that performers might be subjected to changing physical limitations as they age, and that such limitations reflect directly on various technical categories in their performances such as vibrato, intonation or sound production.[9]

Our phylogenetic tree displays no overall affinities regarding traditional partitions into performance schools. As mentioned in the introduction, such classification should be regarded as highly artificial and even irrelevant, as the typical course of study for twentieth-century performers involves many different teachers throughout their years of training. The standard route taken by most violinists analysed in our work involved several major teachers from different backgrounds (this can be seen in Table 5 in Appendix B). It should be noted, however, that on small, local scales (such as the relation between siblings or between 'uncle' and 'nephew' performances) the school affiliation criterion does seem to hold some degree of influence. This fact is interesting, as it may lead us to the conclusion that while there are no common unifying qualities that define one performance school as opposed to another, in certain contexts performance schools may yet play some part.

As presented earlier (see introduction), several musicological studies discuss the homogeneity in the execution of central musical parameters found among HIP. This observation, while not directly contended by our tree, is questioned herewith: while HIP performances are all grouped in the same, 'youngster' subtree, they are not clustered as closely together as one might expect. Interestingly, three performances located in this subtree, which are not explicitly classified as 'historically-informed', may be linked to other aspects of 'stylistic awareness'. Both Telmanyi and Gähler are exponents of the curved bow tradition,[10] while Milstein's 1975 recording has been found in a recent study to be highly influenced by certain aspects of HIP performance practice (see Fabian & Ornoy, 2009).

[6]In his discussion of changing performance styles throughout the years, Wilkinson notes that '... on the whole most recorded musicians for whom we have a lifetime's output seem to have developed a personal style early in their career and to have stuck with it fairly closely for the rest of their lives' (see Leech-Wilkinson, 2009b, p. 250).

[7]For studies examining stylistic change of performance patterns traced among twentieth-century prominent composers see Lebrecht (1990) (addressing Mahler's recordings), Cook (2003) (on Stravinsky), and Park (2009) (on Prokofiev). For studies pointing to performers' individual change of style over the course of time see Katz (2003) (examining multiple recordings made of the same piece by Kreisler, Menuhin and others), Leech-Wilkinson (2009a) (on Arleen Auger, Fischer-Dieskau, Kreisler, and others), and Fabian and Ornoy (2009) (on Heifetz and Miltstein).

[8]As earlier presented, such observations (except, perhaps, the manner of vibrato execution), agree with previous findings, which have used these two recordings to detect Milstein's style change over the years. See Fabian and Ornoy (2009).

[9]Addressing Joachim's apparent poor use of vibrato displayed in his recordings, Styra Avins has pointed to the artists' old age as being 'a particularly severe enemy of vibrato'. See Avins (2003, p. 28).

[10]The 'Vega' (also 'Bach') bow has a round shape and an easily maneuvered mechanism of hair tautness that enables the simultaneous projection of a multiple-stop chord. Its use, presented by its supporters as the one used by Bach and as best suited for his violin music, has never gained real popularity amongst violinists save a few (see Schweitzer, 1950; Boyden, 1965; Spivakovsky, 1967; Schroeder, 1970; Haylock, 2000; Sartorius, 2008).

All in all, it seems that no single category may be decisively sensitive to HIP or non-HIP performance style: It seems that over the years the influence of HIP performance aspects has grown, affecting the attitudes of performers not formally associated with the HIP agenda (such as Tetzlaff or Zehetmair as an example). This may explain why HIP performances are not clustered together but appear scattered across the 'young' subtree. More interestingly, the rather recent performance of Kremer (recorded in 2005) may indicate how widespread and pervasive the HIP approach has become, influencing not only young up and coming performers, but also acclaimed senior performers who were educated in an era and school much different in approach.

Some results displayed in the tree are unexpected. For instance, the pairing of Suk and Ehnes is quite surprising, as these two performers clearly come from vastly different backgrounds. The pairing of Hugget and Telmanyi is also somewhat peculiar, as, putting shared school affiliations aside, they also come from distinctly separate backgrounds. Another surprising element is the pairings one would expect, which are not present, such as the pairing of Mintz and Perlman, who belong to the same generation and have a rather similar upbringing. Observing our validation metric for these three pairs indicates that they are problematic in the sense that most performance analysis categories we used are 'indifferent' to them (seven in the case of Mintz and Perlman, eight in the case of Suk and Ehnes, and nine in the case of Hugget and Telmanyi). A category that is 'indifferent' to a certain pair is basically undecided regarding how each performance in the pair should be placed with respect to the other. This means the final position of these pairs is harder to explain compared to other, more obvious pairs. It would appear that the quartet reconciliation methods we employed may indeed be sensitive to such cases of poorly resolved pairwise relationships. Naturally, a pair most categories are indifferent to may be paired together or apart merely due to the constraints imposed by completely different performances alone.

On the other hand, re-examining the pairwise distances induced by each category does shed light on the unexpected proximity of both Suk and Ehnes, whose performances are, judging by the raw data, surprisingly similar (considering the bowing, the chord ratio and the double stop versus arppegio categories, to name a few). To substantiate this observation, we calculated the Pearson correlation of the two 'combined' raw input vectors (all 87 measurements) for Ehnes' and Suk's respective performances, which is 0.63. In order to put this in context, all the pairwise correlations between raw input vectors were calculated ($\binom{29}{2} = 406$ pair correlations all in all). The average correlation is 0.07, with the maximal value being 0.94 and the minimal value being -0.55. The correlation between the Suk and Ehnes performances is the fifth best correlation of the 406 calculated. We note that among the top five best correlations, there are two sibling pairs in the tree (the two Heifetz performances, ranking first with a correlation of 0.94, and the Suk and Ehnes performances), and two pairs of performances in relatively close proximity (the two Kuijken performances, ranking third with a correlation of 0.66, and the performances by Telmany and Gahler, ranked second with a correlation of 0.89). Only one pair of these five is placed relatively far apart in the tree—the performances by Kremer and Suk, ranked fourth with a correlation of 0.65. It is important to note that raw correlation gives more weight to categories with a higher number of measurements (e.g. the double stop versus arpeggio category, which consists of 30 measurements out of the 87 taken). This observation explains, for instance, why the correlation between Telmanyi and Gahler is so high (as both use a historic Vega bow). Our method compensates for this over-representation, as each category has the same weight in the final tree construction, regardless of its number of measurements.

Similarly, while the performances by Perlman and Mintz are not that far off, they are not particularly close either, judging by most categories (the correlation between the raw input vectors for these two performances is 0.36, which is not particularly strong). All in all, it seems that while this is an issue which requires further attention, the results can be justified even when considering such problematic pairings.

To conclude, the findings presented in our work illustrate a compound picture regarding the proximity relations among performances. Several background factors were found to be influential in shaping an artist's approach to performance. It would appear, however, that an attempt to predict generic classifications of performance styles based solely on shared biographical identities holds little promise: the clustering of performances by performers of different backgrounds highlights joint idiosyncratic peculiarities, which seem to overshadow general categorizations. Each performance is an amalgamation of myriad features, primarily based on performers' complex weighing and balancing of the various performance factors at their disposal. Empirical studies aiming to attribute interpretation profiles to 'objective' means such as biographical background or teacher–student influences may prove quite limited. The performance elements examined in this work are not exhaustive: various idiomatic features (such as specific fingering or different bowing techniques, to name a few) are almost impossible to obtain reliably from audio recordings. To these, one should add the numerous factors related to the creative dimension of the recording process itself: performance features mediated and manipulated by producers and engineers (such as dynamics), limitations and restrictions connected to recording technologies (bearing in mind the untrustworthiness of commercial transfers or even original discs with regard to

timbre analysis), performers' psychological state, in, as well as outside, the studio, or even our listening habits as researchers (for a comprehensive discussion of the subject see Cook et al., 2009).

What this work seems to suggest is a new way of understanding the complex interactive process among performers. An algorithmic approach to musical performance analysis provides tools that might shed future light on fundamental aspects of musical performance, such as the very concept of style and its developments, the origin and nature of performance conventions or the ever-lasting mutual relations between originality, idiosyncrasy and particularization to uniformity and general trends. As such, it provides a new outlook on the history of music performance, thus proving valuable in advancing our understanding of musical performances. To the best of our knowledge, this is the first work taking into account such a large variety of performance aspects, and attempting to amalgamate them into a single, unified view by applying computational means. We expect that additional efforts along these lines will refine and improve on our approach.

Acknowledgements

We would like to thank Sagi Snir, for supplying us with his code implementation of the Rao-Snir quartet max-cut heuristic; Dorottya Fabian, for her contribution to collecting the recordings taken for analysis in this paper and for her useful comments; Meinard Mueller, for useful discussions and suggestions; and Jonathan Miller, for his wise comments and help in editing this paper. Many thanks to Matan Gavish for enabling us access to the Stanford Libraries.

References

Adams III, E.N. (1972). Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology*, *21*(4), 390–397.

Almansa, J., & Delicado, P. (2009). Analysing musical performance through functional data analysis: Rhythmic structure in Schumann's Traumerei. *Connection Science*, *21*(2 & 3), 207–225.

Avins, S. (2003). Performing Brahms's music: Clues from his letters. In M. Musgrave & B.D. Sherman (Eds.), *Performing Brahms: Early Evidence of Performance Style* (pp. 11–47). Cambridge: Cambridge University Press.

Ben-Dor, A., Chor, B., Graur, D., Ophir, R., & Pelleg, D. (1998). Constructing phylogenies from quartets: Elucidation of Eutherian superordinal relationships. *Journal of Computational Biology*, *5*(3), 377–390.

Beran, J., & Mazzola, G. (1999). Analyzing musical structure and performance—a statistical approach. *Statistical Science*, *14*(1), 47–79.

Bomar, M. (1987). *Baroque and Gallant elements of the partitas by J.S. Bach* (PhD dissertation). Colorado University, Colorado, USA.

Bowen, J. (1996). Tempo, duration, and flexibility: Techniques in the analysis of performance. *Journal of Musicological Research*, *16*, 111–156.

Bowen, J. (2005). *Bibliography of performance analysis*. Retrieved from http://www.josebowen.com/ bibliography.html

Boyden, D. (1965). *The History of Violin Playing from its Origins to 1761 and its Relationship to the Violin and Violin Music*. London: Oxford University Press.

Boyden, D. (1980). Violin: Technique, since 1785. In S. Sadie (Ed.), *The New Grove Dictionary of Music and Musicians* (Vol. 19, pp. 839–840). Oxford: Oxford University Press.

Brown, P. (1997). *Sempre Libera: Changes and variation in singing style in Verdi's La Traviata from the first century of operatic sound recordings* (PhD dissertation). University of New South Wales, Sydney, Australia.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In F.R. Hodson, D.G. Kendall, & P. Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences* (pp. 387–395). Edinburgh: Edinburgh University Press, Edinburgh.

Butt, J. (2002). *Playing with History*. Cambridge: Cambridge University Press.

Chor, B. (1998). From quartets to phylogenetic trees. In *SOFSEM' 98: Theory and Practice of Informatics* (Lecture Notes in Computer Science, Vol. 1521/1998, pp. 36–53, doi: 10.1007/3-540-49477-4_3). Berlin: Springer.

Cohen, D. (1969). Patterns and frameworks of intonation. *Journal of Music Theory*, *13*, 66–91.

Cohen, D., & Katz, R. (1968). Remarks concerning the use of the Melograph in ethnomusicological studies. *Yuval*, *1*, 155–164.

Cook, N. (2003). Stravinsky conducts Stravinsky. In J. Cross (Ed.), *The Cambridge Companion to Stravinsky* (pp. 175–191). Cambridge: Cambridge University Press.

Cook, N., Clarke, E., Leech-Wilkinson, D., & Rink, J. (Eds.). (2009). *The Cambridge Companion to Recorded Music*. Cambridge: Cambridge University Press.

Cseszko, F. (2000). *A comparative study of Joseph Szigeti's 1931, Arthur Grumiaux's 1960, and Serjiu Luca's 1983 recordings of Johann Sebastian Bach's sonata no. 1 in G minor for solo violin* (DMA dissertation). University of Wisconsin-Madison, USA.

Dahlback, K. (1958). *New Methods in Vocal Folk Music Research*. Oslo: Oslo University Press.

Dart, T. (1954). *The Interpretation of Music*. London: Hutchinson University Library.

Day, T. (2000). *A Century of Recorded Music: Listening to Musical History*. New Haven, CT/London: Yale University Press.

Dreyfus, L. (1983). Early music defended against its devotees: A theory of historical performance in the twentieth century. *The Musical Quarterly*, *69*, 297–322.

Efrati, R.R. (1979). *Treatise on the Execution and Interpretation of the Sonatas and Partitas for Solo Violin and the Suites for Solo Cello by Johann Sebastian Bach*. Zürich and Freiburg: Atlantis.

Fabian, D. (2003). *Bach Performance Practice 1945–1975: A Comprehensive Review of Sound Recordings and Literature*. Aldershot: Ashgate.

Fabian, D. (2005). Towards a performance history of Bach's Sonatas and Partitas for Solo Violin: Preliminary investigations. In L. Vikárius & V. Lampert (Eds.), *Essays in Honor of László Somfai: Studies in the Sources and the Interpretation of Music* (pp. 87–109). Lanham, MD: Scarecrow Press.

Fabian, D., & Ornoy, E. (2009). Identity in violin playing on records: Interpretation profiles in recordings of solo Bach by early 20th-century violinists. *Performance Practice Review*, *14*. Retrieved from http://scholarship.claremont.edu/cgi/viewcontent.cgi?article=1232&context=ppr

Fabian Somorjay, D. (2000). Musicology and performance practice: In search of a historical style with Bach recordings. *Studia Musicologica*, *41*, 77–106.

Field, E.I. (1999). *Performing solo Bach: An examination of the evolution of performance traditions of Bach's unaccompanied violin sonatas from 1802 to the present* (PhD dissertation). Cornell University, Cornell, USA.

Golomb, U. (2005). Rhetoric and gesture in performances of the First Kyrie from Bach's Mass in B minor (BWV 232). *Journal of Music and Meaning*, *3*(4). Retrieved from http://musicandmeaning.net/issues/showArticle.php?artID=3.4

Halász, P. (1995). The Hungarian violin school in the context of Hungarian music history. *Hungarian Music Quarterly*, *6*, 13–17.

Haskell, H. (1988). *The Early Music Revival—A History*. London: Thames & Hudson.

Haylock, J. (2000). Playground for angels. *The Strad*, *111*, 726–733.

Haynes, B. (2007). *The End of Early Music: A Period Performer's History of Music for the Twenty-First Century*. Oxford: Oxford University Press.

Hellaby, J. (2009). *Reading Musical Interpretation: Case Studies in Solo Piano Performance*. Burlington: Ashgate.

Huson, D., & Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267.

Jiang, T., Kearney, P.E., & Li, M. (2000). A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application. *SIAM Journal of Computing*, *30*(6), 1942–1961.

Katz, M. (2003). Beethoven in the age of mechanical reproduction: The violin concerto on record. *Beethoven Forum*, *10*, 38–55.

Katz, M. (2004). *Capturing Sound: How Technology has Changed Music*. Berkeley: University of California Press.

Katz, M. (2006). Portamento and the phonograph effect. *Journal of Musicological Research*, *25*, 211–232.

Kenyon, N. (Ed.). (1988). *Authenticity and Early Music—A Symposium*. Oxford: Oxford University Press.

Lankovsky, M. (2009). *The pedagogy of Yuri Yankelevich and the Moscow Violin School, including a translation of Yankelevich's article 'On the initial positioning of the violinist'* (DMA dissertation). City University of New York, New York, USA.

Lauer, T. (1997). *Categories of national violin schools* (DMA dissertation). Indiana University, Bloomington, USA.

Lawson, C., & Stowell, R. (1999). *The Historical Performance of Music: An Introduction*. Cambridge: Cambridge University Press.

Lebrecht, N. (1990). The variability of Mahler's performances. *Musical Times*, *131*, 302–304.

Leech-Wilkinson, D. (2009a). *The Changing Sound of Music: Approaches to Studying Recorded Musical Performance* (online). London: The Centre for the History and Analysis of Recorded Music (CHARM). Retrieved from http://www.charm.rhul.ac.uk/studies/chapters/intro.html

Leech-Wilkinson, D. (2009b). Recording and histories of performance style. In N. Cook, E. Clarke, D. Leech-Wilkinson, & J. Rink (Eds.), *The Cambridge Companion to Recorded Music* (pp. 246–262). Cambridge: Cambridge University Press.

Leech-Wilkinson, D. (2010). Performance style in Elena Gerhardt's Schubert song recordings. *Musicae Scientiae*, *14*, 57–84.

Lester, J. (1999). *Bach's Works for Solo Violin—Style, Structure, Performance*. Oxford: Oxford University Press.

Lisboa, T., Williamon, A. Zicari, A., & Einholzer, H. (2005). Mastery through imitation: A preliminary study. *Musicae Scientiae*, *19*, 75–110.

List, G. (1974). The reliability of transcription. *Ethnomusicology*, *18*, 353–377.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, In *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability* (Vol. 1, pp. 281–297). Berkeley: University of California Press.

Madsen, T., & Widmer, G. (2006). Exploring pianist performance styles with evolutionary string matching. *International Journal of Artificial Intelligence Tools*, *15*(4), 495–513.

Margush, T., & McMorris, F.R. (1981). Consensus n-trees. *Bulletin of Mathematical Biology*, *43*, 239–244.

Milsom, D. (2003). *Theory and Practice in Late Nineteenth-century Violin Performance: An Examination of Style in Performance, 1850–1900*. Aldershot: Ashgate.

Molina-Solana, M., Lluís Arcos, J., & Gomez, E. (2008). Using expressive trends for identifying violin performances. In *ISMIR 2008*, Philadelphia, PA, USA, Session 4b. Retrieved from http://www.ugr.es/~miguelmolina/publications/molina-ismir08.pdf

Moore, M. (1974). The Seeger Melograph model C. *Selected Reports in Ethnomusiclogy*, *1*, 3–13.

Musgrave, G., & Sherman, B. (Eds.). (2003). *Performing Brahms: Early Evidence of Performance Style*. Cambridge: Cambridge University Press.

Nannestad, G.B. (2004). Respecting the sound—from aural event to ear stimulous. *Audio Engineering Society 116th Convention*, Berlin, Germany, pp. 1–9.

Ornoy, E. (2006). Between theory and practice: Comparative study of early music performances. *Early Music*, *34*(2), 233–249.

Ornoy, E. (2007). An empirical study of intonation in performances of J.S. Bach's Sarabandes: Temperament, 'melodic charge' and 'melodic intonation'. *Orbis Musicae*, *14*, 37–76.

Ornoy, E. (2008). Recording analysis of J.S. Bach's G Minor Adagio for solo violin (excerpt): A case study. *Journal of Music and Meaning*, *6*. Retrieved from http://musicandmeaning.net/issues/showArticle.php?artID=6.2

Park, J.H. (2009). *A performer's perspective: A performance history and analysis of Sergei Prokofiev's 'Ten piano pieces', op. 12* (DMA dissertation). Boston University, Boston, MA, USA.

Philip, R. (1992). *Early Recordings and Musical Styles: Changing Tastes in Instrumental Performance, 1900–1950*. Cambridge: Cambridge University Press.

Philip, R. (2004). *Performing Music in the Age of Recording*. London: Yale University Press.

Repp, B. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America*, *92*(5), 2546–2568.

Rink, J. (2001). The line of argument in Chopin's E minor Prelude. *Early Music*, *29*(3), 435–444.

Rink, J. (Ed.). (2002). *Musical Performance—A Guide to Understanding*. Cambridge: Cambridge University Press.

Robinson, D.F., & Foulds, L.R., Comparison of phylogenetic trees. *Mathematical Bioscience*, *53*(1–2), 131–147.

Sapp, C.S. (2007). Comparative analysis of multiple musical performances. In *Proceedings of 8th International Conference on Music Information Retrieval, (ISMIR 2007)*, Vienna, Austria, pp. 497–500.

Sapp, C.S. (2008). Hybrid numeric/rank similarity metrics for musical performance analysis. In *Proceedings of 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Vienna, Austria, pp. 501–506.

Sartorius, M. (2008). *The Baroque German Violin Bow*. Retrieved from http://www.baroquemusic.org/barvlnbo. html

Schroeder, R. (1970). *Ob's wohl am Bogen liegt? Geigererinnerungen aus sieben Jahrzehnten* (PDF file). Property of Bogenforschungsgesellschaft e.V., Sankt Augustin.

Schwarz, B. (1977). The Russian School transplanted to America. *Journal of the Violin Society of America*, *3*, 27–33.

Schweitzer, A. (1950). Reconstructing the Bach violin bow. *Musical America*, *70*, 5–13.

Seashore, C.E. (1938). *Psychology in Music*. New York: McGraw-Hill.

Seeger, C. (1951). An instantaneous music notator. *Journal of the International Folk Music Council*, *III*, 103–106.

Sevier, Z.D. (1981). Bach's solo violin sonatas and partitas: The first century and a half. *Bach—The Quarterly Journal of the Riemenschneider Bach Institute*, *12*, 11–19, 21–29.

Snir, S., & Rao, S. (2006). Using max cut to enhance rooted trees consistency. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *3*(4), 323–333.

Spivakovsky, T. (1967). Polyphony in Bach's works for solo violin. *The Music Review*, *28*, 277–288.

Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtress. *Journal of Classication*, *9*(1), 91–116.

Strimmer, K., & von Haeseler A., (1996). Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, *13*(7), 964–969.

Taruskin, R. (1995). *Text & Act—Essays on Music and Performance*. Oxford: Oxford University Press.

Timmers, R. (2007). Vocal expression in recorded performances of Schubert songs. *Musicae Scientiae*, *11*(2), 237–268.

Trapani, J., & Richter, D. (1985, October). Signal processing through dynamic equalization. *Audio Engineering Society 79th Convention*, New York, USA, pp. 1–13.

Tresize, S. (2009). The recorded document: Interpretation and discography. In N. Cook, E. Clarke, D. Leech-Wilkinson, & J. Rink (Eds.), *The Cambridge Companion to Recorded Music* (pp. 186–209). Cambridge: Cambridge University Press.

Turner, R. (2004). *Style and tradition in string quartet performance: A study of 32 recordings of Beethoven's Op. 131 Quartet* (PhD dissertaion). University of Sheffield, Sheffield, UK.

Weiss, A.S. (1992). Review of Harvith, J. and Edwards, S. (1987). Edison, musicians, and the phonograph: A century in retrospect. *SubStance*, *212*, 127–131.

Williams, H.M. (1931). Experimental studies in the use of the tonoscope. *Psychological Monographs*, *41*(4), 266–327.

Appendix A: Quartets

A quartet is an unrooted subtree with four leaves. For every choice of four leaves there are three quartet topologies on these leaves. Given a set of quartets over leaves, finding a tree that is consistent with as many of them as possible is a hard computational task (Steel, 1992). There are efficient heuristics, such as quartet puzzling, quartets max-cut and many others (see Strimmer & von Haeseler, 1996; Ben-Dor et al., 1998; Jiang et al., 2000; Snir & Rao, 2006; to mention a few), which produce a tree that typically exhibits a good agreement with the input quartets. Such a heuristic—the Rao-Snir quartet max cut—is employed in our trees construction.

Figure 13 depicts a concrete example of an (unknown) tree with five leaves, 1 through 5. Out of the five possible quartets ($\binom{5}{4} = 5$), we are given four quartets topologies, which in this case are inconsistent due to possibly corrupt data (no tree is compatible with all of them). The heuristic we apply reconstructs a tree compatible with three of these four quartet topologies, displayed in Figure 14.

Appendix B: Performance schools

Table 5 and Figure 15 contain concentrated data regarding the educational background and performance school affiliation of all performers discussed in this paper.

Table 5. Concentrated information regarding the performance school of each of the performers.

Performer's name	Performers' school affiliation
Szigeti	Hubay's pupil ('Hungarian school')
Heifetz	Auer's pupil ('Russian school'-St. Petersburg)
Enescu	Hellmesberger's pupil ('Viennese school'), Marsick's pupil ('Parisian school')
Telmányi	Hubay's pupil ('Hungarian school')
Milstein	Stoliarsky's pupil ('Russian school'-Odessa), Auer's pupil ('Russian school'-St. Petersburg)
Menuhin	Enescu's pupil ('Viennese school'), Persinger's pupil ('Franco-Belgian school' + 'American school')
Grumiaux	Enescu's pupil ('Parisian school' + 'Vienna school')
Szeryng	Flesch's pupil ('German school'), Frenkel's pupil ('Russian school'-St. Petersburg)
Suk	(Sevcik)/Kocian pupil ('Czech school')
Végh	Hubai's pupil ('Hungarian school')
Luca	Rostal's pupil ('German school'), Galamian's pupil ('American school'), affiliated with 'HIP school'
Ricci	Persinger's pupil ('Franco-Belgian school' + 'American school')
Zehetmair	Rostal's pupil ('German school'), Milstein's pupil ('Russian school'-Odessa)
Kuijken	Raskin's pupil ('Franco-Belgian school'), affiliated with 'HIP school'
Mintz	Feher's pupil ('Hungarian school'), Stern's pupil ('American school')
Perlman	Galamian's DeLay's pupil ('American school')
Tetzlaf	Levin's pupil ('American school'), Haiberg's pupil ('German school')
Hugget	Kuijken's pupil ('Franco-Belgian school'), Parikian's pupil ('Hungarian school'), affiliated with 'HIP school'
van Dael	Goldberg's pupil ('German school'), affiliated with 'HIP school'
Wallfisch	Grinke's pupil ('Franco-Belgian school' + 'English school'), affiliated with 'HIP school'
Gahler	Schroeder's pupil, Brero's pupil ('German school')
Podger	Comberti's pupil (affiliated with 'HIP school')
Ehnes	(Galamian)/Chaplin's pupil ('American School')
Szenthelyi	Kovacs's pupil ('Hungarian school')
B. Brooks	Goldberg's pupil ('German school')
Kremer	Oistrach's pupil ('Russian school'-Moscow)

Appendix C: Methods (extended version)

In this appendix we fully describe, in a greater level of technical detail, the data collection and processing phases, and the quartet based approach we used for constructing the phylogenetic trees. Musicological considerations, correlation results and the validation process are not discussed in this appendix, as they were already fully reviewed in Section 2.

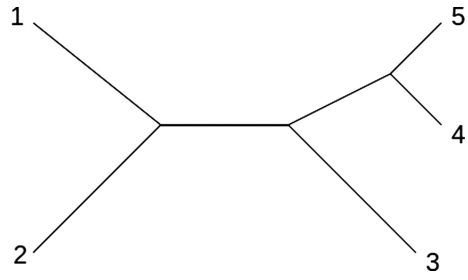

Fig. 13. Original tree.

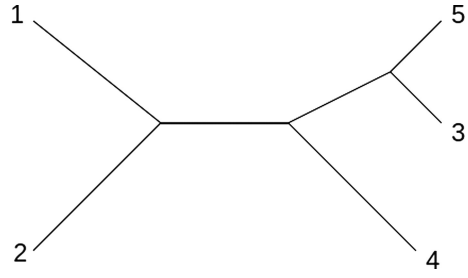

Fig. 14. Tree resolved from input quartets. Explanation: suppose we have some abstract original tree of structure ((1,2), (3), (4,5)) (see Figure 13). We are not provided with this tree, but with some quartets which purportedly originate from it. The input data is corrupt, however, so in reality we obtained three 'true' quartets and one 'false' quartet: (1,2—3,4) (1,2—3,5) (2,3—4,5) (1,4—3,5). We have no means of knowing which of the quartets is false, but we may immediately notice that in this specific case, no tree can satisfy all our input quartets. The best we could hope for is a tree which satisfies three out of the four input quartets. This may not necessarily be the original tree: for instance, in this case, the tree ((1,2), 4, (3,5)) satisfies three out of four quartets (the same as the original one).

Data

29 performances of Bach's sonatas and partitas for solo violin were collected. Segments of two specific movements (BVW 1001 Adagio and BVW 1005 Adagio) were selected for data analysis. For each performance, measurements belonging to 10 distinct categories were examined. In each category, a number of features were extracted (between 1 and 30 per category), adding up to an 87-dimensional measurements vector per performance. The 10 categories are:

- *Bowing*—the marking of bow changes. Ten features (determined by auditory means). Each feature represents whether bow direction was changed, partially changed or unchanged at ten diachronic points in the sampled section of BVW 1001 (Gm adagio). These points were chosen after meticulously studying the recordings, so that at each such point, at least one performer had indeed changed bow direction. Direction change, partial direction change and no change were encoded by the numerical values $[1, 0.5, 0]$, respectively.
- *Chord ratio*—the ratio between the lowest and the highest notes in the sampled chords. Fourteen features (measured with the Sonic Visualiser software package, see the Musical Considerations subsection for a detailed explanation of this category).
- *Double stop/arpeggio*—represents whether the chord is an arpeggio or a double stop—a vector of 30 measurements (one measurement for each chord in the analysis range, determined by auditory means). For each chord in the sequence, performers who utilized a Vega bow scored 2 (no breaking), performers who utilized a double stop scored 1 ('half-break'), and performers who opted for an arpeggio scored 0 (complete break).
- *Count of double stops in C adagio*—one feature (double stop frequency).
- *Vibrato*—split into three 'sub-features': depth, speed and onset. Nine features (3×3 measurements of depth, speed and onset for three sample notes. measured with the Sonic Visualiser software package).
- *Duration per bar*—11 features (11 bars) (measured with the Sonic Visualiser software package). These measurements were taken from the sampled segment of BVW1005 (C adagio), since each bar in this segment is meaningful in terms of phrasing.
- *Tempo changes*—10 features $(10 = 11 - 1)$—the difference between adjacent duration measurements (the differences are based on the duration measurements, which were collected for the previous category). This measure is converted to a $[-1, 0, 1]$ scale:
  - (-1) if bar j is more than 10% slower than bar $(j - 1)$
  - $(+1)$ if bar j is more than 10% faster than bar $(j - 1)$
  - 0 otherwise (i.e., no significant change).

  The choice of 10% as a threshold was essentially empirical. Smaller changes could easily be explained as marginal and unintentional performance inconsistencies (which do not reflect interpretive considerations), and furthermore, are less discernable from measurement noise.

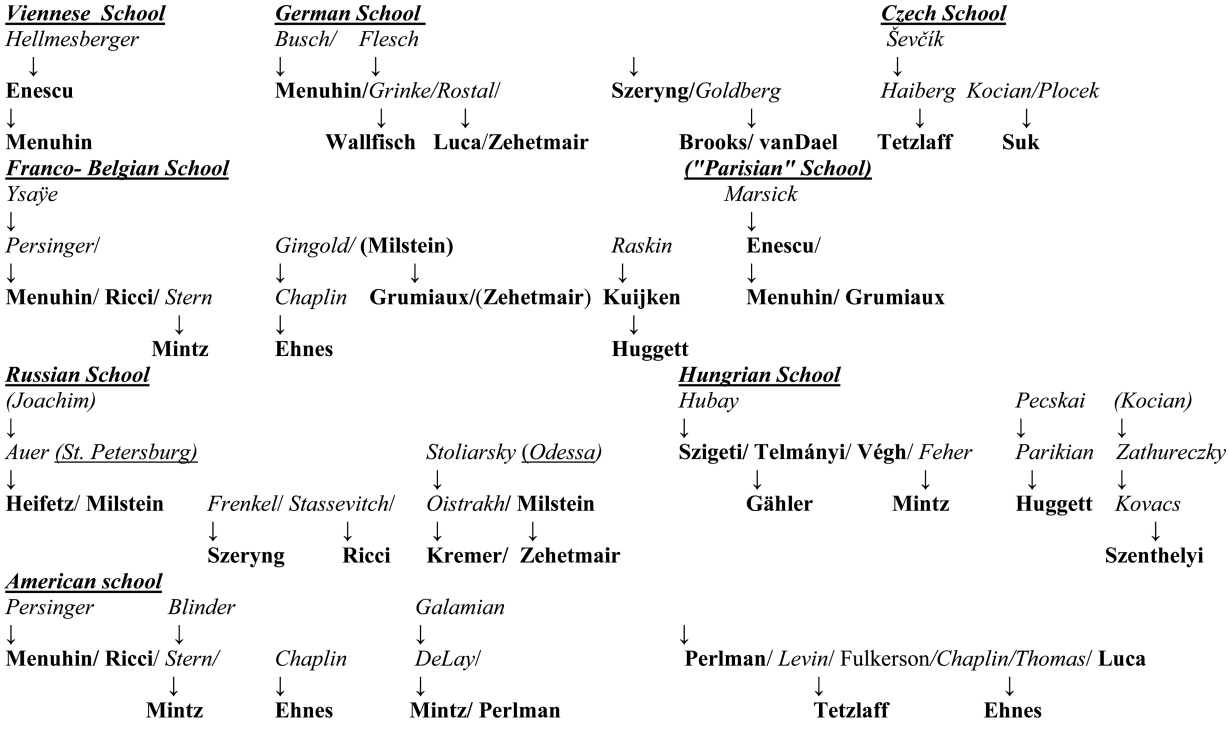

Fig. 15. Division to performance schools and educational background of performers.

- *Total duration*—one feature (measured with the Sonic Visualiser software package).
- *Dotting ratio*—ratio between adjacent long and short notes (measured with the Sonic Visualiser software package, based on the first bar of the CM Adagio movement).
- *Standard deviation of the tempo changes*—the standard deviation of the tempo changes vector (of length 10)—one feature. This feature is useful in quantifying the tempo variance of a given performer.

Standard deviation of tempo changes versus skewness of tempo changes

Our tempo changes criteria represents the trend between adjacent tempo measurements (quantized to three levels, implying some tolerance to noise). Therefore, the standard deviation of this criterion doesn't directly correspond to the variance in performance tempo, but rather the variability of tempo trends in a given performance. Another possible measure for tempo variability, which has been used in music performance analysis (Leech-Wilkinson, 2010), is the skewness. The skewness of a random variable X is defined as follows:

$$\text{skewness} = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right],$$

where E is the expectation operator, μ is the mean and σ is the standard deviation.

The difference between the two measures is that the skewness is not sensitive to the actual frequency of tempo changes. For example, if we consider the series $[+1, -1, +1, -1]$, representing sequential accelerations and decelerations of tempo, the skewness will be 0, same as for the series $[0, 0, 0, 0]$. Our measure, on the other hand, would assign the value 1 to the first, and 0 to the second.

Normalization

Having collected the data, each entry in each category was normalized separately so its average value (over the 29 performances) would be 1. This action, beyond being common practice, is also crucial for information-theoretic reasons, as it aptly quantifies the 'surprise', or weight, of certain actions compared to others. Let us assume we have a one-dimensional binary vector representing whether a performer did or did not apply a certain technique at some point in time. If 9 out of 10 performers applied this measure, than after normalization the weight of this action would be $1/0.9 = 1.111\ldots$ If, however, only 1 performer out of the 10 applied this measure, then after normalization the weight of this action would be $1/0.1 = 10$. This makes sense, because an action taken by most performers is less surprising than an action taken by only a few, and should therefore contribute less to distance considerations.

Processing—quartet construction

The combined data was first examined, and a unified Euclidean distance matrix was calculated. By applying the Buneman (1971) tree criteria on the unified distance matrix, we discovered that it has a highly incongruent nature—the resulting tree was completely unresolved—a star. For this reason, classic distance based approaches (such as neighbour joining and k-means clustering) were deemed inapplicable. We thus decided to adopt a quartet based approach (Ben-Dor et al., 1998; Chor, 1998; Jiang et al., 2000).

Initially, we work with each category separately, choosing those quartet topologies, which have a clear support. For every possible quartet ($\binom{29}{4} = 23,751$ possibilities) each category provided a 'vote' for the strongest possible topology supported by the pairwise Euclidean distances inducted in that category. For every four items, a, b, c, d, there are three possible topologies—$(a, b|c, d)$, $(a, c|b, d)$, and $(a, d|b, c)$. For a given topology, e.g. $(a, b|c, d)$, we say it is consistent with the distances between the items if the sums of the 'diagonals' are approximately equal (i.e. $d(a, c) + d(b, d) \cong (a, d) + d(b, c)$, and the sum of the distances along the 'non-diagonal' $(d(a, b) + d(c, d))$ is considerably smaller than the 'diagonal' sums.

Processing—tree construction

At this stage, each category 'votes' for a certain topology if the two diagonal sums are close enough to one another and distant enough from the non-diagonal sum. Let $sum1 = d(a, b) + d(c, d)$, $sum2 = d(a, c) + d(b, d)$, $sum3 = d(a, d) + d(b, c)$, and assuming (WLOG) they are sorted so that $sum1 \leq sum2 \leq sum3$. The decision whether a quartet is supported or not is based on two predetermined thresholds: $parameter1$, representing the minimal requirement for the ratio between $sum1$ and $sum2$ (which should be considerably smaller than 1), and $parameter2$, representing how much the ratio between $sum2$ and $sum3$ may deviate from 1 (should be close to 1).

Formally the requirements can be simply stated as:

$$sum1 \leq parameter1 * sum2$$

$$sum3 \leq (1 + parameter2) * sum2.$$

If no topology satisfies the predetermined thresholds for a given quartet, the category 'abstains' w.r.t. these four performances (i.e. prefers no topology).

After the voting process is complete, we retain only the quartet topologies which have substantial support from the data. These would be quartets with a high enough number of supporting categories for the 'winning' topology (higher than a predefined threshold,

which we shall refer to as *parameter*3), and with a high enough ratio of supporting versus opposing categories (a support rate higher than a predefined threshold, which we shall define as *parameter*4). Since the resultant list of quartet topologies is dependent on the four predetermined parameters we just described (parameters 1 to 4), a vast set of possible parameter configurations was tried. This resulted in 2100 lists of quartet topologies, one per each set of possible parameters.

An unrooted phylogenetic tree was constructed for each list of quartet topologies, using Snir and Rao's (2006) 'quartets max-cut' heuristic (the problem of building a tree from quartets is computationally intractable, thus a heuristic is called for) (Steel, 1992). Subsequently, each tree is given a score, based on its rate of accordance with the list of quartet topologies from which it was constructed. In addition, the size of this 'support list' was also considered, as well as the number of splits the resultant tree contains (we give preference to trees which are based on a large number of quartet topologies, and trees which are resolved enough to display meaningful information). Different scoring functions (using these inputs as their arguments) were attempted. Eventually, after removing all the trees constructed by lists with less than 2500 quartets, and trees with less than 15 splits, we remained with 224 'meaningful' trees. These 224 trees were sorted by the following scoring function:

$$\begin{aligned} \textit{Tree score} = {} & 2 * (\textit{agreement rate}) + \textit{splits rate} \\ & + \textit{support rate} \end{aligned}$$

with

$$\begin{aligned} \textit{agreement rate} = {} & \textit{fraction of quartets topologies in} \\ & \textit{the support list satisfied by the tree} \end{aligned}$$

$$\textit{splits rate} = \frac{(\textit{number of splits in the tree})}{(\textit{max possible number of splits})}$$

$$\textit{support rate} = \frac{\textit{size of support list}}{\textit{max number of quartet topologies}}$$

(max possible number of splits = 28 and Max number of quartet topologies = 23,751).

Processing—consensus trees

A list of consensus trees (majority vote, see Margush and McMorris, 1981) was constructed—for the 20, 40, 60, 80 and 100 highest scoring trees (out of the list of 224 meaningful trees we described earlier). The topologies of the five resultant consensus trees were in relatively high proximity with one another, as can be seen in Table 6, presenting the pairwise distance between the trees (according to the Robinson–Foulds metric). In this sense, we can say that the resultant consensus trees are 'stable'.

In order to compare the various consensus trees, we devised two 'tree quality' measures. The first is to take the aforementioned 224 quartet lists, which gave rise to the meaningful trees, and calculate the average agreement rate between these lists and the consensus tree. A second measure is to use this set of quartet lists to calculate the average extent of support (calculated as $\textit{number of supporting quartets} - (\textit{num of opposing quartets} + \varepsilon)$) for siblings in the tree. That is, to calculate the average rate of support for all sibling pairs with respect to each quartet list, and then calculate the overall average of this measure. The logic behind this measure is that ideally, if the data is consistent, and two performances which are indeed 'true siblings' are paired together, no quartet topology should place these two performances on opposite sides, and thus the support rate for this pair should be extremely high (basically it is $\textit{number of quartets in list} - (0 + \varepsilon)$) across all lists. Since the different categories are not consistent, we cannot expect such support, and in fact we often encounter striking inconsistencies (for almost all pairs of performances, at least one category is in favour of placing the two performances together, and one is in favour of placing the performances apart). Generally speaking, the more distant a pair of performances is, the lower their rate of support should be (and indeed a strong correlation between these two factors was found, as discussed in the validation section). The total score for the consensus trees was simply the product of the two measures described above. The scores for the five consensus trees constructed are presented in Table 7.

In addition to these five consensus trees, we also calculated our scores with respect to 10 randomly

Table 6. Robinson–Foulds distances of the five resulting trees (*the maximal distance score in our case is* 2 * 28 = 56).

	Cons_20Trees	Cons_40Trees	Cons_60Trees	Cons_80Trees	Cons_100Trees
Cons_20Trees	0	–	–	–	–
Cons_40Trees	8	0	–	–	–
Cons_60Trees	8	0	0	–	–
Cons_80Trees	14	6	6	0	–
Cons_100Trees	10	2	2	4	0

generated trees (randomly selected binary tree topology, and randomly placed leaves). Furthermore, we calculated our scores for the star tree on these 29 performances at the leaves. For random binary trees we would expect 1/3 of the quartets to be satisfied, on average. In addition, we would expect the support ratio for siblings to be low, because the neighbourhood relations are independent of the data. Indeed our results fit these expectations. For the star tree, the quartet agreement rate is by definition 0, and we would expect the sibling support to be $(1/3)/(2/3) = 0.5$, which is indeed observed (the outcome is not exactly 0.5 since as previously mentioned, in our calculations we add $\varepsilon = 10^{-4}$ to the opposition count to avoid the possibility of division by zero). The selected tree was cons_80Trees (the tree constructed via consensus over the 80 highest scoring trees out of the 224 meaningful ones). The second best tree was cons_20Trees (the tree constructed over the 20 highest scoring trees). We note that cons_80Trees obtained better results than cons_20Trees for $\sim 60\%$ of the quartet lists considering the first measure, and $\sim 80\%$ considering the second measure, thus indicating the resultant tree is consistently superior to the 'competing' option.

Table 7. Scores for the five consensus trees, 10 random trees, and the star tree (*last two rows calculated as 'sanity checks'*).

Consensus tree	Average agreement rate	Support ratio for siblings	Total score
cons_20Trees	0.839856853	2.908976416	2.443123777
cons_40Trees	0.844628076	2.457206168	2.075425318
cons_60Trees	0.844628076	2.457206168	2.075425318
cons_80Trees	**0.844234085**	**2.995664472**	**2.529042054**
cons_100Trees	0.845905116	2.457206168	2.078563269
avg over 10 random trees	0.31	0.067	0.021
star tree	0	0.4995	0

Appendix D: Additional clustering analysis examples

The following figures display the clustering results for the double stop versus arpeggio and the chord ratio categories. These analyses were made twice—with Gaehler and Telmanyi (Figures 16 and 17), who are clear outliers for obvious reasons (both performers use the non-standard Vega bow), and without them (Figures 18 and 19). We note that removing outliers affects both the clustering results and the principal components used for the visualization of multivariate categories.

We observe that the clustering divisions made by each category are very different, and that no obvious clusters may be discerned. However, the proximity relations revealed in some of the cases are quite telling (for instance, Kuijken83, Brooks, Luca, Mintz84 and Wallfisch are well grouped, and so are the performances Kuijken99, Milstein75, Podger, Teztlaff and Hugget).

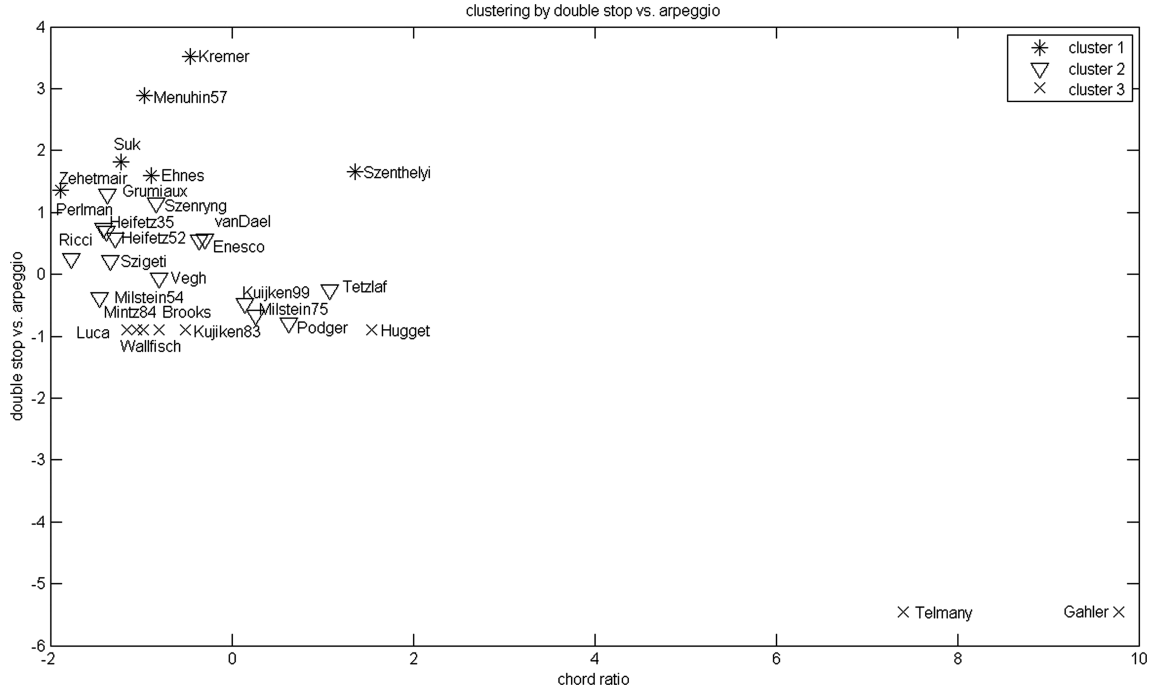

Fig. 16. Double stop/arpeggio versus chord ratio, clustering according to double stop/arpeggio.

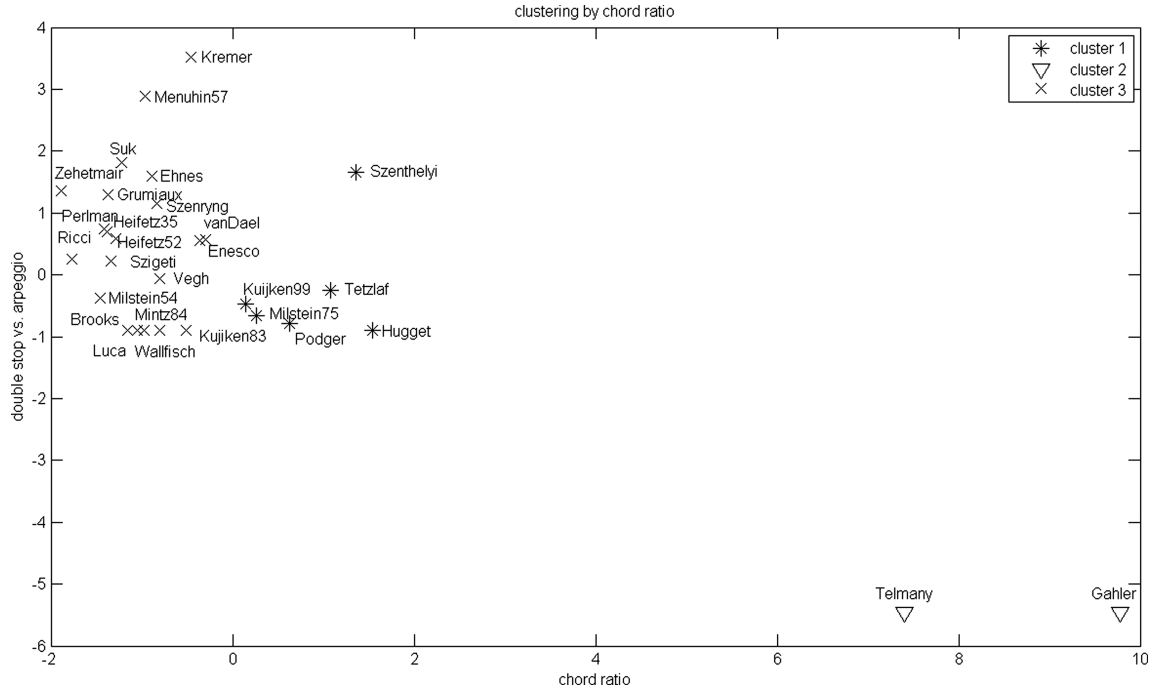

Fig. 17. Double stop/arpeggio versus chord ratio, clustering according to chord ratio.

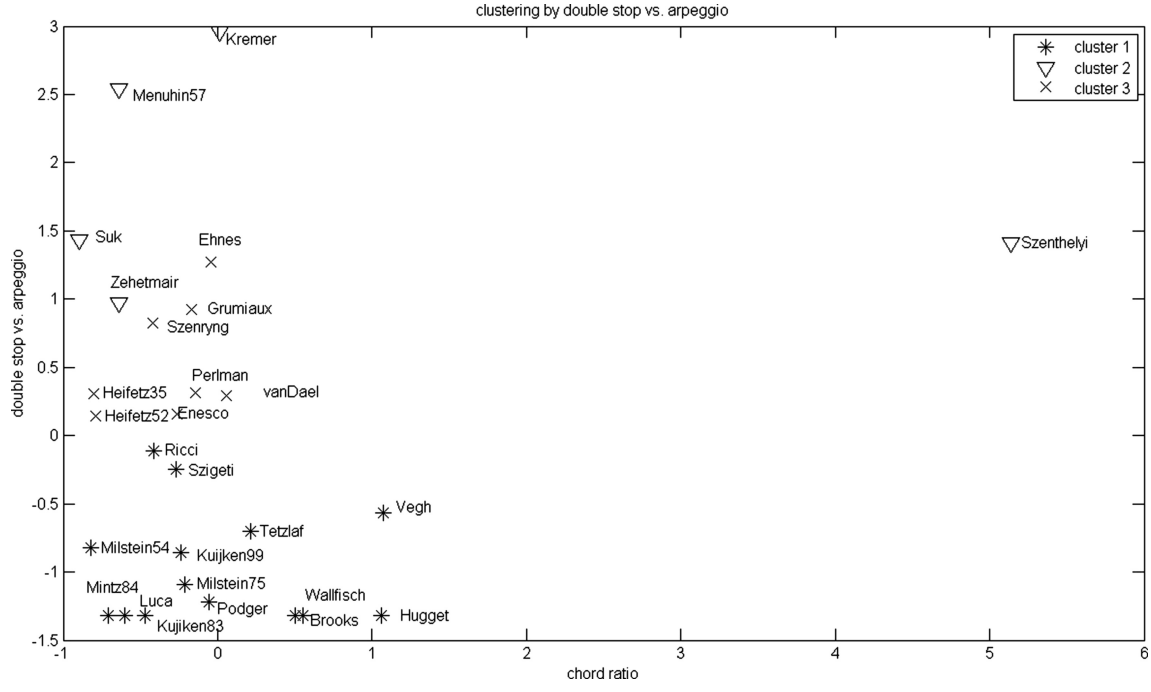

Fig. 18. Double stop/arpeggio versus chord ratio, clustering according to double stop/arpeggio (without Telmanyi or Gahler).

The following figures present the clustering results using the bowing and the mid-phrase durations categories. Again, we repeat the process twice—with a clear outlier performance—Enesco (Figures 20 and 21), and without it (Figures 22 and 23).

Once again, no clear clustering results could be discerned. We note that it is hard to draw clear conclusions from the bowing category alone, as it is relatively homogenous. One may also observe that the two Heifetz performances are similarly distinct in terms of bowing, as are the performances by Menuhin and Wallfisch (perhaps surprisingly). The mid-phrase durations category is even harder to analyse in terms of proximity relations, although it may be noted that the

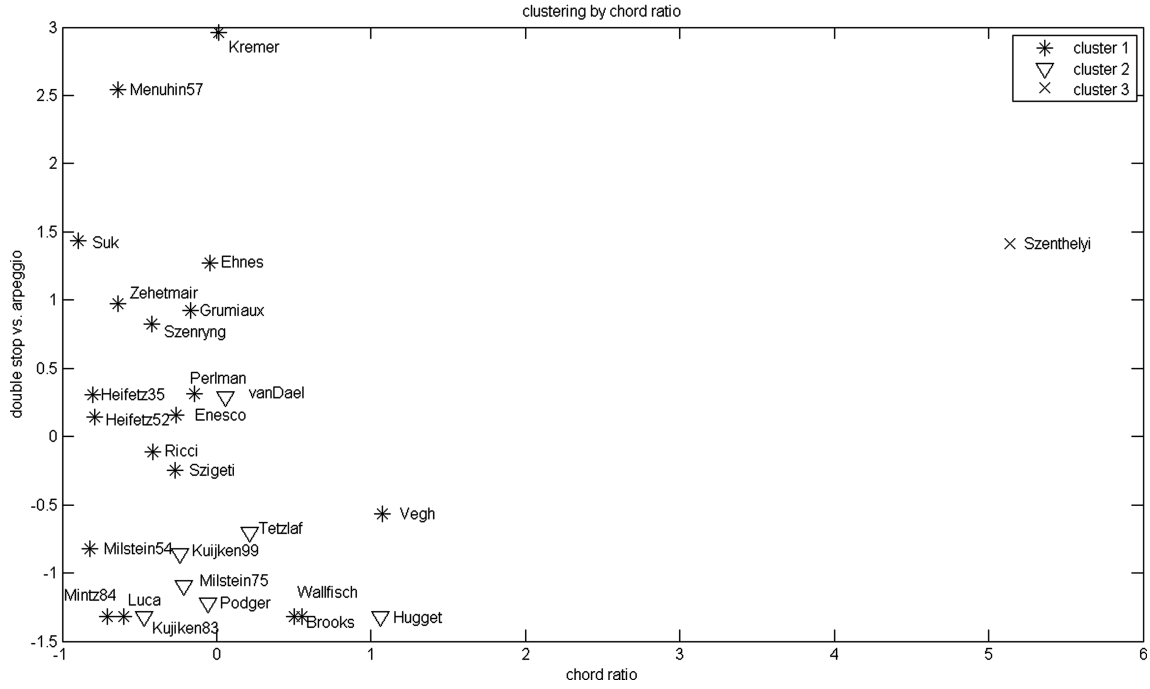

Fig. 19. Double stop/arpeggio versus chord ratio, clustering according to chord ratio (without Telmanyi or Gahler).

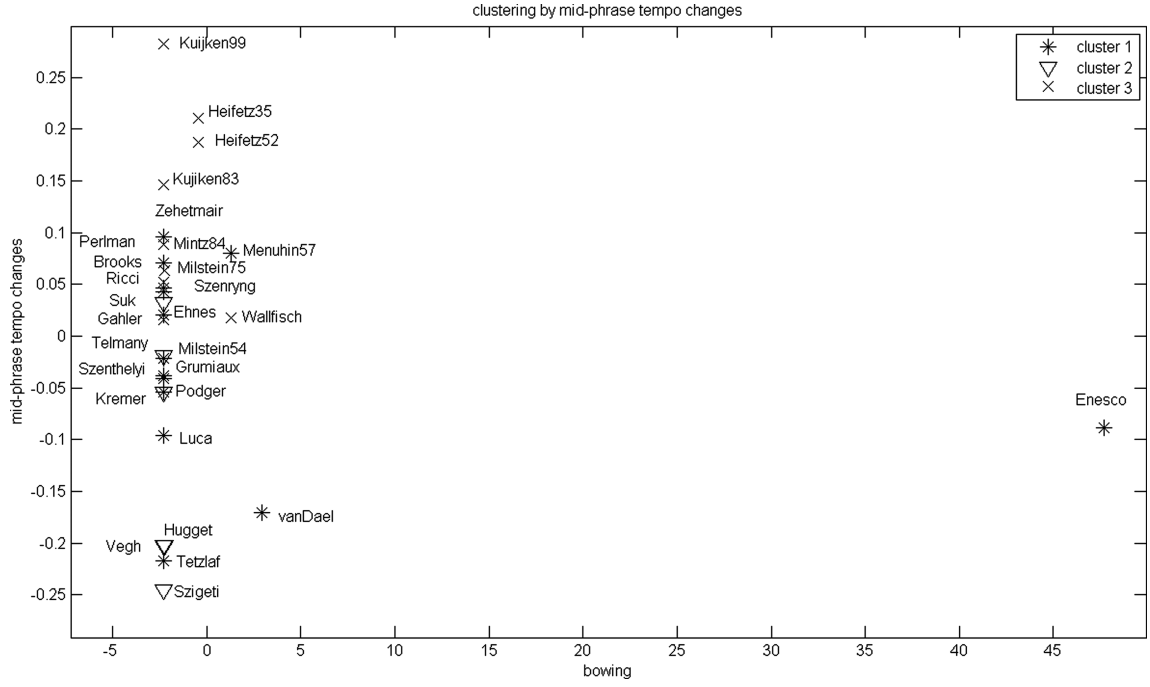

Fig. 20. Mid-phrase durations versus bowing, clustering according to mid-phrase durations (with Enesco).

performances by Vegh, Hugget, Tetzlaff, and Szigeti are fairly well grouped by that category.

All in all, considering the clustering results, it is clear that not only each category implies wholly different categories. No category alone is enough to make general meaningful observations regarding the complex relations between the 29 performances analysed. For this reason, our quartet based voting mechanism was devised, meant to reconcile the inherent differences between categories and allow for more reliable observations regarding the interconnections between the performances.

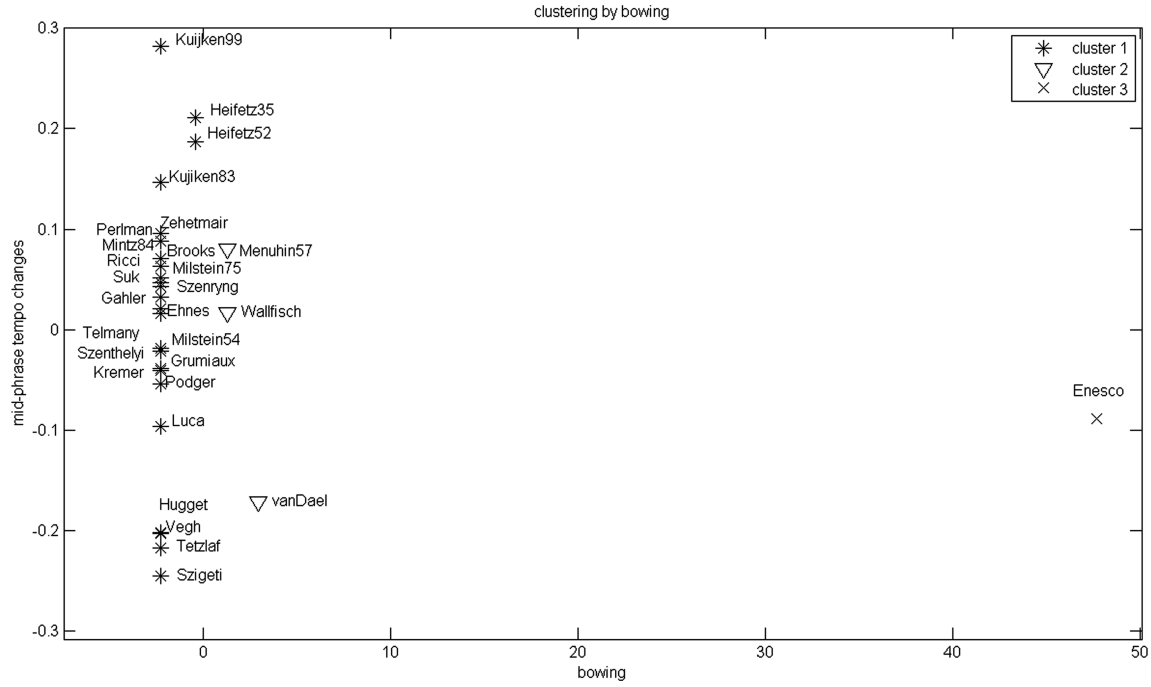

Fig. 21. Mid-phrase durations versus bowing, clustering according to bowing (with Enesco).

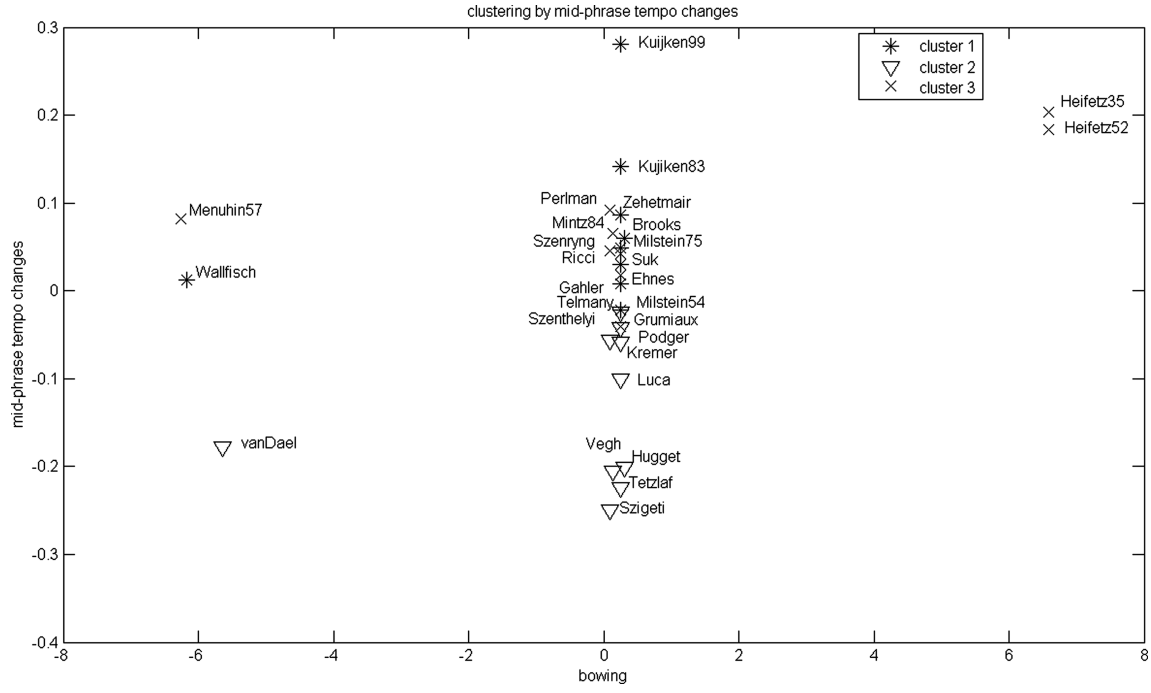

Fig. 22. Mid-phrase durations versus bowing, clustering according to mid-phrase durations (without Enesco).

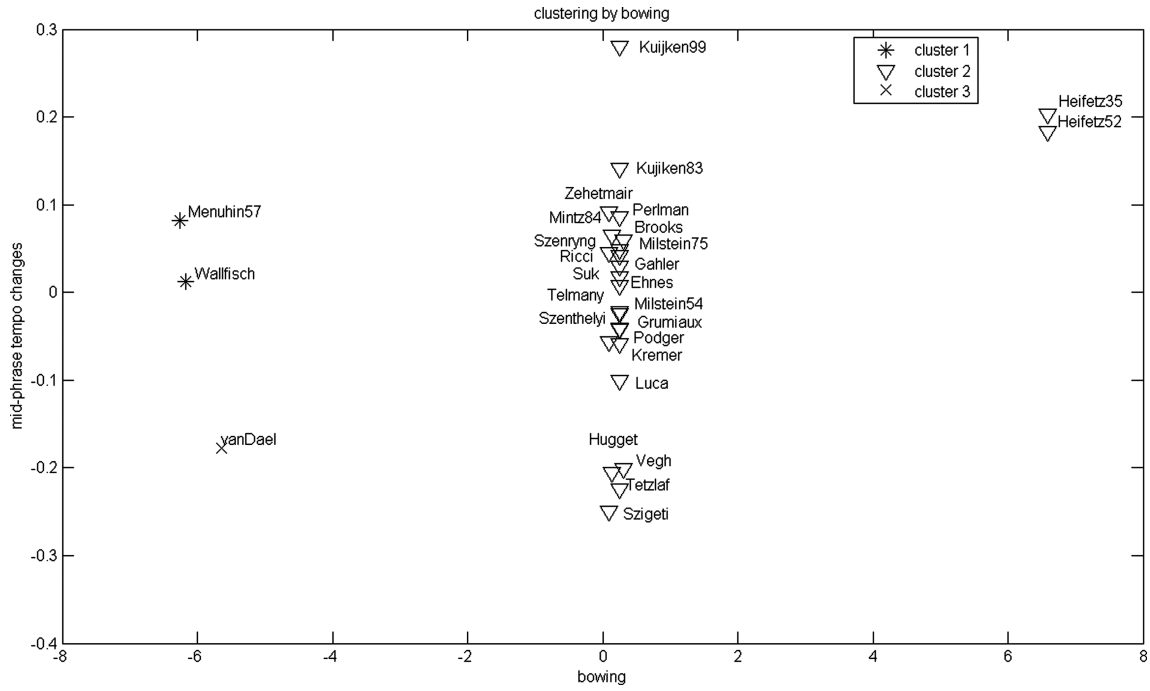

Fig. 23. Mid-phrase durations versus bowing, clustering according to mid-phrase durations (without Enesco).

Appendix E: List of performances

Table 8 details performers' names, date of birth and date of recording.

Table 8. Performers' names, date of birth and date of recording.

Performer's name	Date of Birth	Recording date
Enescu	1881	1948
Szigeti	1892	1931
Telmányi	1892	1954
Heifetz	1901	1935
Heifetz	1901	1952
Milstein	1903	1954
Milstein	1903	1975
Végh	1905	1971
Menuhin	1916	1957
Szeryng	1918	1968
Ricci	1918	1981
Grumiaux	1921	1960–1961
Suk	1929	1971
Gahler	1941	1998
Luca	1943	1977
Kuijken	1944	1983
Kuijken	1944	2001
Perlman	1945	1986
van Dael	1946	1996
Kremer	1947	2005
Szenthelyi	1952	2002
Wallfisch	1952	1997
Hugget	1953	1995
Mintz	1957	1984
B. Brooks	1959	2003
Zehetmair	1961	1983
Tetzlaff	1966	1994
Podger	1968	1999
Ehnes	1976	1999–2000