Symbolic Melodic Similarity: State of the Art and Future Challenges

Abstract

Fostered by the introduction of the Music Information Retrieval Evaluation eXchange (MIREX) competition, the number of systems which calculate Symbolic Melodic Similarity has recently increased considerably. In order to understand the state of the art, we provide a comparative analysis of existing algorithms. The analysis is based on eight criteria that help characterising the systems, and highlighting strengths and weaknesses. We also propose a taxonomy which classifies algorithms based on their approach. Both taxonomy and criteria are fruitfully exploited for providing input for new forthcoming research in the area.

«Start article»

The advent of the Internet has made a large quantity of audio and symbolic musical data freely available. The analysis of these data can provide useful insights into several aspects of music. By comparing many musical pieces, it is possible to abstract relevant rules and processes which characterise a particular style. Also, the analysis of large databases can improve our understanding of the generative process, shedding light on the evolutionary path undergone by music over time. In order to capitalise upon the significant body of knowledge currently stored within online music datasets, a number of reliable and efficient automatic tools have been developed over the last decades. Melodic similarity-detection algorithms are an instance of such tools. When used on

online musical datasets, they can provide valuable information on intra- and inter-work melodic relationships and on the underlying melodic structures of the pieces analysed.

Given two or more sequences of notes, Symbolic Melodic Similarity aims to evaluate their degree of likeness, as human listeners are able to do. This task has relevance both within the academy and in industry. For instance, beyond the purely academic benefits of identifying the degree of likeness between musical pieces and composer-practices afforded by melodic similarity systems, plagiarism detection constitutes an example of a practical application of this task with clear legal and commercial implications. Many algorithms for judging melodic similarity have been introduced over the years. Even though such tools perform essentially the same task, they may be based on theories and methods which belong to radically different disciplines. For example, there are some algorithms based on principles from music theory, others based on cognitive constraints, and others which implement notions from pure mathematics.

Thanks to the "Symbolic Melodic Similarity" track of the MIREX competition (Downie 2008), introduced in 2005, the number of tools in this field has increased dramatically. However, the last published surveys on Symbolic Melodic Similarity consider algorithms developed up to 2004 (Müllensiefen and Frieler 2006; Hofmann-Engl 2010). This evident lack of evaluation of state-of-the-art techniques is the first motivation for this paper. Accessibility is a second motivation for the paper. The literature on Symbolic Melodic Similarity is distributed across many different sources, which cover numerous topics from computer science to music theory. This survey brings together recent studies on Symbolic Melodic Similarity, describing them in a concise way, so that researchers can form an initial overview of the approaches used by other scholars The paper proposes a highly modular taxonomy which allows the effective categorisation of techniques according to the approach they exploit. We identify eight criteria, which summarise the most relevant aspects of each melodic similarity

algorithm. By analysing the state of the art, we are also able to provide guidelines and recommendations for a further development of the field. Moreover, the aforementioned taxonomy and the criteria facilitate the classification and comparison of future systems.

The paper is structured as follows. Firstly, we outline relevant background information and related works. Secondly, we outline the identified criteria. Next, the taxonomy is presented and the considered systems are briefly described and categorised. Then, algorithms are compared with regard to the eight criteria. Finally, by synthesising the knowledge gained with the analysis, we propose new directions for research and we offer conclusions.

Background

Stephen Downey defined Music Information Retrieval (MIR) as "a multidisciplinary research endeavour that strives to develop innovative content-based searching schemes, novel interfaces, and evolving networked delivery mechanisms in an effort to make the world's vast store of music accessible to all" (Downie 2004). The interdisciplinary environment of MIR encompasses many fields, such as computer science, psychology, musicology, music cognition and signal processing. MIR combines these disciplines in order to create real-world applications which are capable of extracting relevant information from music. MIR techniques have been applied to solve a large number of tasks such as music recommendation, automatic music transcription, and track separation. MIR systems can represent music in two ways: audio and symbolic. Systems which adopt audio representation directly encode musical information through digital audio formats such as WAV and MP3. Applications which use symbolic representation are usually based on MIDI and MusicXML formats. Symbolic encoding allows the system to manipulate musical items without recourse to signal processing, while having a clear representation of the overall score.

Symbolic Melodic Similarity is a central issue of MIR. Many applications exploit musical similarity in order to retrieve pieces from a database, to perform musical analysis, and to categorise music. Essentially, all systems which are based on melodic similarity try to find musical utterances which match the information needed by the user, expressed in a query. Melodic similarity is also used in technologies that are revolutionising the way people experience music. Applications which evaluate melodic similarity improve the accessibility of musical databases, allowing users efficiently to retrieve the musical information they need. Furthermore, such systems can enhance the understanding of the structure of music itself. Indeed, musicologists can exploit applications to track stylistic traits of musical pieces, and to trace the occurrences of musical patterns within and between musical works. This can deepen our understanding of musical style, while actively promoting a new quantitative/empirical analytical approach among musicologists and music theorists.

A major issue with melodic similarity is that assessing the likeness between two musical phrases is an extremely difficult process, which reflects the cognitive complexity of the task. Most of the complexity associated with melodic similarity detection arises from the multidisciplinary nature of this process. Melodic similarity spans several elements of music theory, ethnomusicology, cognitive science and computer science, all of which have to be considered simultaneously. Music theory suggests how to identify syntactically relevant musical structures. Ethnomusicology accounts for the variety and the cultural dependency of melodies from distinct geographical regions. Music cognition describes the basic cognitive processes which humans deploy in order to recognise melodies as similar. Finally, computer science affords a means to create "intelligent" computational systems able to embed the insights provided by the other fields. Although melodic similarity has been extensively investigated in all of these disciplines, there is no agreed, clear-cut definition of the field yet. Even scholars from the

same background disagree on the ambit and methodologies of melodic similarity.

This survey focuses on MIR systems performing melodic similarity analysis which have been presented since 2004. Systems introduced before have already been reviewed (Müllensiefen and Frieler 2006; Hofmann-Engl 2010). For the sake of completeness, we briefly report the main strategies developed before 2004, which strongly affected the later research environment. In 1996, McNab, Smith, Witten, Henderson, and Cunningham (1996) introduced an algorithm based on the edit distance between motives. Cambouropoulos (1998) devised a system based on melodic contrast that dynamically creates motivic categories containing similar melodic structures. In the same year, Maidín (1998) developed a system which exploits the distance in pitch between two melodies, weighted through correlation and difference coefficients. Later, Downie (1999) presented a system which assesses melodic similarity based on N-grams. Finally, Meek and Birmingham (2002) developed an algorithm which exploits hidden Markov chains.

In 2005, MIREX (the Music Information Retrieval Evaluation Exchange) was established (Downie 2008). This organization runs an important annual competition which aims to compare state-of-the-art algorithms and systems relevant for MIR. One of the several tracks of the contest is Symbolic Melodic Similarity and in this respect MIREX helps to facilitate cross-fertilisation among researchers, while constantly fostering the improvement of the techniques and strategies adopted in melodic similarity research. MIREX has become the main forum for researchers and practitioners interested in evaluating and comparing algorithms which span several tasks of MIR. As a consequence, many systems considered in this article have been tested or trained on MIREX benchmarks. Indeed, this competition has been providing the MIR community with a large set of useful benchmarks for almost ten years.

Criteria

We have formulated eight criteria that are useful for evaluating melodic similarity systems. They have been designed to investigate systems' functionality from a number of different perspectives: flexibility, similarity evaluation, training and validation.

- *Polyphony*. This indicates the ability of a system to deal with musical pieces that include one or more voices. Monophonic systems can evaluate melodic similarity only between pieces containing a single melodic line.

- *Scope*. This indicates the musical genres and styles a system is able to investigate. Some methods have a general scope, being able to analyse a wide range of musical styles. Others evaluate melodic similarity only in a specific type of music (e.g., folk, classical, pop).

- *Similarity Function*. In order to calculate melodic similarity, algorithms rely on functions which provide a quantitative measure. Usually, such functions are based on well known geometrical, mathematical, cognitive or musical notions. This criterion indicates which methodology has been used by a specific system.

- *Musical Parameters*. Melodic similarity is a multidimensional problem. In order to evaluate the similarity between melodies, tools should consider several musical parameters, such as pitch, duration, etc. This criterion lists the musical parameters that have been taken into account by a system.

- *Musical Representation*. The way in which musical pieces are represented can vary considerably between systems. Each representation shows strengths and weaknesses that affect the operation of the algorithm. Pieces have variously been encoded as strings, numbers, trees or graphs.

- *Experiment-Based*. Melodic similarity is deeply rooted in perception and cognition. To develop efficient algorithms, it is sometimes necessary to rely draw upon experiments in human perception. The experiment-based criterion shows whether or not the similarity function of a system has been designed based upon the results of some cognitive empirical research.

- *Training*. This criterion indicates whether or not an algorithm has been trained – using some machine-learning techniques – on some specific dataset. Trained systems are expected to have a good performance on musical pieces that are comparable to those used for training.

- *Empirical Validation*. Algorithms exploit a similarity function for obtaining a quantitative value of likeness between two musical pieces. This criterion investigates whether or not the similarity function has been validated.

Taxonomy

In this section, we briefly describe the 15 algorithms for melodic similarity detection considered. The systems are organised into four categories, according to the strategies they exploit, namely: *cognition*, *music theory*, *mathematics* and *hybrid*. Since the melodic similarity task is intrinsically interdisciplinary, researchers have addressed it by exploiting techniques which derive from very different areas. This taxonomy reflects the four main areas of investigation adopted in Music Information Retrieval.

It is worth noting that some of these disciplines overlap, thus it can sometimes be hard unequivocally to classify techniques based on approaches at the edge of two areas. For example, several theories of music such as pitch-class set theory (Forte 1973) and the Generative Theory of Tonal Music (GTTM) (Lerdahl and Jackendoff 1985) are strongly

founded on mathematics. Therefore, any algorithm built upon one of these theories could be safely classified both as a member of the music theory and the mathematics categories. In such cases, and for the sake of clarity, we have classified systems according to the category which qualitatively fits them better.

This taxonomy aims to provide a first step in the direction of a comprehensive ontology for melodic similarity techniques. The proposed framework can be extended straightforwardly, by adding new categories as well as by identifying meaningful subcategories. Indeed, additional categories might be needed when new approaches are developed.

Systems Based on Cognition

Cognitive constraints have only recently been used for evaluating melodic similarity. Algorithms which rely on this approach are usually based on one or both of two methods: the linear combination of cognition-based metrics and human-tailored pattern recognition. The work of Roig, Tardón, Barbancho, and Barbancho (2013) and Vempala and Russo (2015) exploits the linear combination of different metrics, based on the evaluation of differences between pairs of musical features extracted from the melodies, such as pitch distance, pitch direction, and rhythmic salience. On the other hand, de Carvalho Junior and Batista (2012) proposes a system – one based on a form of Prediction by Partial Matching (Cleary and Witten 1984) – that simulates the operation of the human auditory cortex in assessing the similarity between given MIDI files.

Systems Based on Music Theory

Music theory has long provided tools for understanding the structure of melodies. Therefore, several melodic similarity tools are built upon theories such as the GTTM

(Lerdahl and Jackendoff 1985), the Implication/Realisation (I/R) Model of Narmour (Narmour 1992) and Schenkerian Analysis (Forte and Gilbert 1982).

Grachten, Arcos, Lopez de Mantaras et al. (2004) proposes a melodic similarity measure based on the I/R model. Specifically, the algorithm tries to implement most of the innate processes presented by the I/R model. Melodies are annotated by performing I/R analysis. Annotations are then used for comparing different melodies. For overcoming one major issue of the I/R model – the impossibility of unambiguously differentiating the intervallic direction of a sequence of notes – Yazawa, Hasegawa, Kanamori, and Hamanaka (2013) designed an algorithm based on an extended I/R model, one that includes a number of new symbols for improving expressivity.

In a Schenkerian vein, Orio and Rodà (2009) introduces an approach based on the representation of musical pieces as hierarchical graphs, in order to identify the relative structural importance of notes. The most relevant notes in a melody are then compared in order to find similarities between melodic segments.

Systems Based on Mathematics

Systems in this category use a number of mathematical approaches which serve as a basis for evaluating the degree of likeness between melodies. Many of these algorithms represent musical data as functions within an abstract space and exploit notions from geometry as a means of comparison. Other common mathematical strategies are based on statistical analysis or information-retrieval techniques.

Aloupis, Fevens, Langerman, Matsui, Mesa, Nuñez, Rappaport, and Toussaint (2006) presents two algorithms that exploit a geometrical approach. Melodies are represented as polygonal chains within a pitch-time abstract plane, and their similarity

is calculated as the minimum area between polygonal chains. In 2010, the S2 and W2 algorithms were introduced (Lemström 2010; Laitinen and Lemström 2010). Both are geometric algorithms that work with point-set representation of music, and are designed to process polyphonic music. The similarity between the query pattern and the target melody is defined by the number of elements that match between them, after the application of some invariants. Finally, Urbano (2013) proposes three different systems: ShapeH, ShapeTime, and Time. All three rely on a geometric model that encodes melodies as curves in a pitch/time space, that are then compared.

Bohak and Marolt (2009) investigates how melody-based features relate to folk-song variants. Specifically, the authors extract 94 melody-based features from each melody, that are then used for performing comparisons.

Wolkowicz and Kešelj (2011) proposes six different algorithms (WK1–6) that exploit text information retrieval approaches. They extract features from an input dataset of MIDI files. String-based methods can be directly applied to such input files, since none of the notes is concurrent or overlapping. Then, they build N-grams, i.e., substrings of N consecutive tokens (i.e., notes), which are used to calculate similarity between melodies. In the framework proposed by Frieler (2006), melodies are represented by series of arbitrary length in an abstract space of events. N-grams are then used for measuring the similarity of two melodies. N-grams have an identity which allows one to assess their degree of similarity, thus providing further information.

Hybrid Systems

In order to maximise the efficiency of algorithms, hybrid systems combine several techniques which usually belong to two or more of the aforementioned categories. The SIMILE toolbox (Müllensiefen and Frieler 2004; Frieler and Müllensiefen 2005) is based

on a linear combination of *c.* 50 algorithms which cover different aspects of the Symbolic Melodic Similarity task. Authors gathered rating data by human experts, which were considered as the ground-truth for evaluating and optimising the linear combination of the considered techniques' similarity values.

Fanimae (Suyoto and Uitdenbogerd 2010) combines two similarity measurements: the NGR5 (Suyoto and Uitdenbogerd 2005) approach, which exploits 5-grams for matching melodies by considering only the pitch; and the newly introduced PIOI technique, which evaluates similarity by considering either pitch or duration.

Rizo and Inesta (2010) introduced the UA systems, four methods designed with the objective of obtaining a good trade-off between accuracy and processing time. Two of them are based on tree representation, one on the quantised point-pattern representation, and one is the ensemble of all the previously mentioned methods.

Comparison of Systems

By analysing according to the criteria previously introduced in the corresponding section, it is possible to infer general trends and exceptions in how researchers conceived, designed and implemented their systems. A detailed mapping of each system against our criteria is provided in Figure 1. Each criterion gives an insight into one specific aspect of a melodic-similarity algorithm. Rather than listing decisions taken by authors with regard to each criterion, we outline trends and relevant deviations from trends that have been highlighted by the criteria.

The majority of the systems calculate melodic similarity between monophonic sequences only. However, the system developed by Laitinen and Lemström (Lemström 2010; Laitinen and Lemström 2010), as well as the algorithm built by Suyoto and

	Category	Polyphony	Scope	Similarity function	Musical parameters	Musical representation	Based on Experiments	Trained	Empirical validation
Carvalho Junior, Batista (2012)	cognition	no	general	match frequent sequences of pitch intervals, duration ratios between melodies	pitch intervals, duration ratio	string of symbols	no	no	musical incipits
Roig et al. (2013)	cognition	no	general	linear combination of metrics	downbeat onset, passing note onset, pitch direction, pitch intervals, transposition	not specified	no	no	not specified
Vempala, Russo (2015)	cognition	no	tonal melodies	linear combination of metrics	pitch distance, pitch direction, duration, contour, tonal stability	not specified	yes	yes	tonal melodies
Grachten et al. (2004)	music theory	–	general	edit distance	pitch, contour	string of symbols (I/R symbols, note sequences)	no	no	jazz songs
Orio, Rodà (2009)	music theory	–	music based on harmony	shortest path between two nodes in a graph	harmony, metre, pitch	graph	no	no	musical incipits (RISM)
Yazawa et al. (2013)	music theory	no	general	matched N-gram sequences	pitch direction, duration	string of symbols (extended I/R symbols)	yes	no	folk songs (Essen collection)
Aloupis et al. (2006)	mathematics	no	general	minimum area between polygonal chains	pitch, duration	geometric (polygonal chains)	no	no	not specified
Frieler (2006)	mathematics	–	general	linear combination of N-gram similarity measures	pitch, duration	sequence of symbols	no	no	no
Bohak, Marolt (2009)	mathematics	no	folk songs	statistical differences	melodic complexity, meter, entropy, pitch, duration	96 statistical melodic features	no	yes	folk songs
Laitinen, Lemström (2010)	mathematics	yes	general	number of matching elements between melodies	pitch, duration	point-set representation	no	no	musical incipits
Wolkowicz, Kešelj (2011)	mathematics	no	general	text retrieval methods	pitch	string of symbols	no	yes	musical incipits
Urbano (2013)	mathematics	no	general	change in shape of curves	pitch, duration	geometric (curve in pitch/time plane)	no	no	musical incipits
Frieler, Müllensiefen (2004)	hybrid	no	general	linear combination of metrics (edit distance, N-grams, geometric distance, correlation coefficient)	pitch, duration, contour, tonality, accent structure	not specified	yes	yes	musical incipits (RISM)
Rizo, Iñesta (2010)	hybrid	no	general	linear combination of metrics	pitch, duration	tree	no	no	not specified
Suyoto, Uitdenbogerd (2010)	hybrid	yes	general	linear combination of metrics (N-grams, dynamic programming)	pitch, duration	string of symbols	no	no	musical incipits

Figure 1. Detailed description of systems based on eight identified criteria.

Uitdenbogerd (2010), can additionally evaluate melodic similarity between polyphonic pieces. It should also be noted that some systems do not provide information about their polyphony-analysis capabilities.

Most systems have a general scope, since they can be used to calculate the likeness of any melody belonging to any style. On the other hand, the algorithm developed by Vempala and Russo (2015) focuses on melodies which are rooted in tonality, the system by Orio and Rodà (2009) analyses music in which harmony plays a functional role, and the system built by Bohak and Marolt (2009) works with folk songs only.

There is no dominant trend in similarity functions exploited by algorithms. Combining several similarity measures together, by means of a linear combination, is a popular approach (Vempala and Russo 2015; Frieler 2006; Roig, Tardón, Barbancho, and Barbancho 2013; Rizo and Inesta 2010; Müllensiefen and Frieler 2004; Suyoto and Uitdenbogerd 2010). However, in such cases, the specific measures which contribute to form the linear combination change from system to system. Other systems (Urbano 2013; Aloupis, Fevens, Langerman, Matsui, Mesa, Nuñez, Rappaport, and Toussaint 2006) calculate the differences of shape and area between curves which represent melodies in abstract mathematical spaces. There are also approaches based on statistics (Bohak and Marolt 2009), the shortest path between two nodes in a graph (Orio and Rodà 2009), and text-retrieval methods (Wolkowicz and Kešelj 2011). Also, some traditional methods, such as edit distance (Grachten, Arcos, Lopez de Mantaras et al. 2004) and N-grams (Yazawa, Hasegawa, Kanamori, and Hamanaka 2013), are exploited.

Musical parameters considered for calculating similarity are almost always pitch and duration. Information about time and frequency seem to be expressive enough to allow a precise judgement of likeness between musical excerpts. However, some systems also rely on parameters such as harmony (Orio and Rodà 2009), contour

(Vempala and Russo 2015; Grachten, Arcos, Lopez de Mantaras et al. 2004; Müllensiefen and Frieler 2004), and pitch direction (Vempala and Russo 2015; Yazawa, Hasegawa, Kanamori, and Hamanaka 2013). It should be noted that only the algorithm developed by Bohak and Marolt (2009) includes a number of unusual and interesting parameters, such as melodic complexity, metric accent, and entropy.

The diversity of approaches used to calculate melodic similarity is matched by (perhaps even springs from) the heterogeneity of musical representations A common method of encoding musical information is through sequences of symbols (Frieler 2006; de Carvalho Junior and Batista 2012; Rizo and Inesta 2010; Wolkowicz and Kešelj 2011). Other strategies explored to represent music are trees (Rizo and Inesta 2010), graphs (Orio and Rodà 2009), statistical features (Bohak and Marolt 2009), and I/R symbols (Grachten, Arcos, Lopez de Mantaras et al. 2004; Yazawa, Hasegawa, Kanamori, and Hamanaka 2013). Finally, two algorithms which rely on mathematics (Urbano 2013; Aloupis, Fevens, Langerman, Matsui, Mesa, Nuñez, Rappaport, and Toussaint 2006) represent melodies as curves in an abstract mathematical space.

Unfortunately, the large majority of systems reviewed in this survey are not based on first-hand experiments in music perception. Indeed, only three algorithms (Vempala and Russo 2015; Yazawa, Hasegawa, Kanamori, and Hamanaka 2013; Müllensiefen and Frieler 2004) are guided by direct experimental enquiry. Other algorithms either implement theoretical hypotheses, or follow well-established notions in music perception.

Moreover, only four algorithms out of the fifteen considered have been trained on a set of melodies (Vempala and Russo 2015; Bohak and Marolt 2009; Müllensiefen and Frieler 2004; Wolkowicz and Kešelj 2011). The other systems do not benefit from the use of such machine-learning methods. As empirically observed in many areas of computer

System	MIREX Edition (Participants)				
	2010 (13)	2011 (10)	2012 (6)	2013 (5)	2014 (4)
Carvalho Junior, Batista (2012)	-	-	6	-	-
Roig et al. (2013)	-	-	-	3	-
Yazawa et al. (2013)	-	-	-	4	-
Laitinen, Lemström (2010)	4	-	-	-	-
Wolkowicz, Kešelj (2011)	-	4	-	-	-
Urbano (2013)	1	1	1	1	1
Rizo, Iñesta (2010)	7	-	-	-	-
Suyoto, Uitdenbogerd (2010)	8	-	-	-	-

Table 1. Rankings of the systems in the Melodic Similarity Track of different editions of MIREX. The number of participants is shown in brackets.

science and Artificial Intelligence, machine learning is a promising technique that could enhance the performance of algorithms considerably.

We observe that there is a lack of empirical validation among reviewed approaches. Although most methods have been tested on groups of melodies, the lack of empirical validation for some of the algorithms (Aloupis, Fevens, Langerman, Matsui, Mesa, Nuñez, Rappaport, and Toussaint 2006; Frieler 2006; Rizo and Inesta 2010; Roig, Tardón, Barbancho, and Barbancho 2013) makes their assessment difficult, if not impossible. However, in the case of the method proposed by Frieler (2006), the lack of validation process is due to the theoretical nature of the article, which introduces innovative mathematical notions which the author acknowledges need future evaluation. Among the systems which provide empirical validation, most have been tested on musical incipits (Urbano 2013; Orio and Rodà 2009; Lemström 2010; de Carvalho Junior and Batista 2012; Müllensiefen and Frieler 2004; Suyoto and Uitdenbogerd 2010; Wolkowicz and Kešelj 2011). Other algorithms have been tested with folk songs (Yazawa, Hasegawa, Kanamori, and Hamanaka 2013; Bohak and Marolt 2009), jazz melodies (Grachten, Arcos, Lopez de Mantaras et al. 2004), and tonal melodies (Vempala and Russo 2015).

Although not conclusive, an analysis of the results of the MIREX competitions in the

Symbolic Melodic Similarity track is informative about the strengths and weaknesses of algorithms. It is worth noting that the best algorithm according to one particular MIREX competition is not necessarily the best system in all situations. From Table 1, it is possible to see that the systems proposed by Urbano (2013) have been the most successful, winning five editions of the MIREX competition. To compare the results achieved by the different algorithms proposed we consider their F1-score, which is the harmonic mean of *precision* and *recall* ranging from 0 to 1. The performance of all the algorithms which competed in the 2010 edition of MIREX was close. None of the systems had a F_1-score greater than 0.30: Urbano (2013) $F_1 = 0.30$; Laitinen and Lemström (2010) $F_1 = 0.27$; Suyoto and Uitdenbogerd (2010) $F_1 = 0.26$; and Rizo and Inesta (2010) $F_1 = 0.26$. In the 2011 campaign, there was a significant increment in the performance of the algorithms which doubled their accuracy. Both the systems proposed by Urbano (2013) and Wolkowicz and Kešelj (2011) achieved the same result, with $F_1 = 0.64$. After 2011, improvements have slowed down. So far, the best performance has been achieved by Urbano (2013) in 2014, with $F_1 = 0.77$. It is difficult to compare these systems with those which did not enter the MIREX competitions, since they have not been assessed in accordance with the same, standardised, evaluation procedure. It is worth mentioning, however, that the algorithm introduced by 2015 obtained a remarkable result, with a performance accuracy of 90%.

Although there has been a significant improvement in the methods introduced in the last decade, not all new systems are better than those developed before 2004. This is the case with the algorithm developed by Yazawa et al. (2013), which perhaps pays the price for experimenting with a new strategy. Nonetheless, it is safe to assert that pre-2004 algorithms have been outperformed in all aspects by newer methods.

Regardless of the rankings obtained by the systems in MIREX, it is possible to identify situations in which algorithms based on different methods perform best. In case

of melodies which differ by a few pitches, intervals and rhythms, an approach based on the number of matching elements in the two melodies (e.g., (de Carvalho Junior and Batista 2012; Laitinen and Lemström 2010)) is likely to be the most effective. This situation is common in both folk and classical styles, where the composer introduces variety by altering a few musical elements of a melody (Figure 2, b). In case of melodies which are substantially different, but which occasionally share similar musical fragments, algorithms based on matched N-gram sequences (e.g., (Frieler 2006; Yazawa et al. 2013)) have an edge, because they are able to detect similarities even if only tangential. This scenario is often found in contrapuntal music (e.g., fugues and motets), where it is common that only the head and the tail of a theme are maintained throughout different repetitions (Figure 2, c). Systems based on a linear combination of different metrics and statistical musical differences (e.g., (Roig et al. 2013; Suyoto and Uitdenbogerd 2010; Vempala and Russo 2015)) are most effective when used on melodies that are loosely similar. Belonging in this category are those melodies which share only a similar contour, or in which ornamental notes are intercalated between the fundamental notes (Figure 2, d). This compositional technique is employed in classical music and it is frequently encountered in pieces built around a single melodic idea. Finally, when the relationship between two melodies results from a mixture of the three situations introduced above, the most effective approach is that of Urbano (2013), which calculate differences in the shape of geometric representations of a melody.

Guidelines and Recommendations

The work done by researchers in the last decade has significantly increased the number of algorithms which calculate melodic similarity and it has improved the variety and quality of approaches. Systems are now capable of finding melodies similar to a target sequence of notes, and searching through large databases of melodies in a more

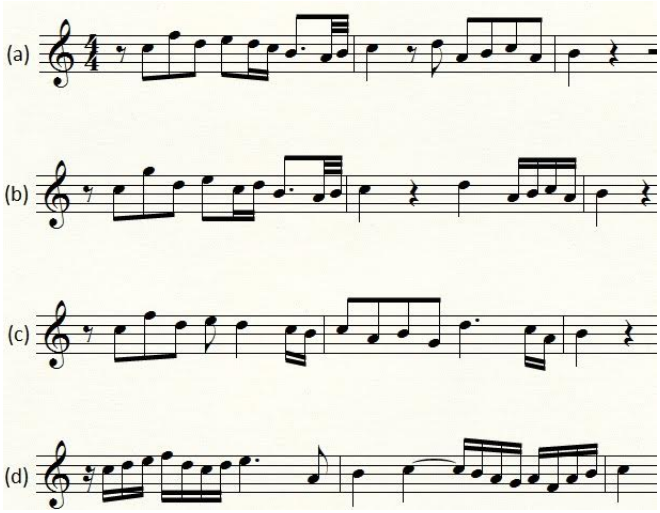

Figure 2. Melodies with different categories of similarity. (a) Model; (b) Melody obtained by changing few pitches and durations of the model; (c) Melody which conserves only the head and the tail of the model; (d) Ornamented version of the model.

reliable, efficient and precise way than before. However, there are still some limitations, and several steps can be taken to improve systems' performance and functionality.

As already mentioned in the Background Section, a clear-cut definition of melodic similarity is missing, As Marsden (2012) points out, the notion of melodic similarity is highly dependent on context. Different models are required to account for similarity judgements carried out by people. Each person considers different sets of melodies. This strongly affects both the way in which systems are evaluated, for instance in the MIREX competition, and the development of new algorithms. The lack of a widely approved definition of melodic similarity is also reflected in the lack of a common ground-truth, shared by all researchers interested in this field, against which the performance of algorithms can be tested. The evaluation framework provided by MIREX is a first attempt to solve this issue. However, as Urbano points out 2013, there are some problems with this music dataset. The MIREX evaluation framework failed to give a consistent measure of performance for the same system over the a number of years. As a consequence, this framework cannot be adopted to benchmark performances of different

systems over time. The solution for a reliable ground-truth for melodic similarity is to have an extremely large database of melodies, in which each melody has a series of scores that indicate the degree of similarity between it and all other melodies in the collection. To develop this evaluation framework, a number of experiments with human listeners are needed. However, since it is necessary to design a large database, these experiments could be too onerous to conduct in a traditional laboratory-based setting. This issue can be overcome by conducting experiments on the Internet, which allows exploitation of the very large number of people available online.

A second issue affecting current systems is that most of them focus on monophonic music only. We believe that next-generation algorithms should be able to analyse polyphonic music and thus find the degree of similarity between two polyphonic excerpts. Polyphonic music in the Western tradition considerably outweighs monophonic music. Therefore, the impact of polyphonic similarity systems on musicology and music theory would be significantly increased. Developing polyphonic similarity analysis tools would also encourage researchers to discover the main differences between evaluating likeness between monophonic and polyphonic musics. This research should be informed by experiments in music perception and, in turn, would provide new insights to musicologists and music theorists.

There is a close relationship between music cognition and melodic similarity algorithms. In order to develop more efficient systems, it is of paramount importance to conduct focused experiments in music perception. The results of these studies could eventually suggest innovative notions and techniques still overlooked in the design of computational systems.

Another relevant point to consider in order to enhance the performance of algorithms is their scope. Most of the methods reviewed have a general scope. In other

words, these systems can be used to analyse melodies belonging to any musical style. However, it is well known that every style has specific rules which help create its unique sound. This is true for melody generation as well. Therefore, if we want to obtain greater efficiency, we need to concentrate on specific styles, rather than developing algorithms able to parse any kind of music. While building these style-specific tools, we might discover some of the deepest rules which contribute to define a style. Understanding these rules would not only improve the efficiency of systems, but would also help musicologists and music theorists understand the complex phenomenon of musical style. Style-specific rules can be extracted, for instance, by using techniques of machine learning. Only a few systems analysed in this survey are based on training strategies. This implicitly indicates that most of the systems developed so far have a general scope, or have been manually configured following developer intuitions. Clearly, training on specific styles will make systems less general. To avoid this pitfall, compound tools, made up of a series of style-specific subsystems, can be designed. In this way, systems would operate distinct subsystems, depending on the style of melodies they are analysing.

A weakness of using machine learning is that there is a trade-off between performance and interpretability (James et al. 2013). The more complex the machine-learning technique employed, the looser the relationship between the predictors (i.e., musical features) and melodic similarity. This phenomenon is due to the potential for overfitting in highly flexible methods. However, if such systems are used only to provide a score of likeness between melodies, the interpretability of the predictive model is not of interest. In this case, the trade-off between performance and interpretability is not a significant concern.

Melodic similarity algorithms based on mathematics have been dominating MIREX competitions for the last five years. In our opinion, this does not mean that algorithms

based on mathematics are intrinsically superior to those based on cognition or on music theory. Rather, this should encourage the improvement of tools based upon different theories, in order to close the performance gap with mathematical systems. Indeed, the task of melodic similarity is highly interdisciplinary and should be tackled from diverse perspectives. For this reason, it would be advisable to develop tools that merge different approaches. By creating hybrid systems, researchers can ameliorate the weaknesses of specific techniques, while augmenting the overall strength of the system.

Conclusions

Melodic similarity systems have a wide range of applications. For instance, they can be used to perform information retrieval on large music datasets, and can help identify music plagiarism. Recently, because of the introduction of the MIREX competition, a large number of algorithms have been developed. In this paper, we have briefly described existing approaches, and we have presented a new modular taxonomy and eight criteria which are instrumental in classifying and comparing melodic similarity algorithms. The taxonomy classifies algorithms based on their approach, i.e., whether they are based on cognition, music theory, mathematics or some hybrid of these. This paper fills the gap between the previous survey in the area (Müllensiefen and Frieler 2006; Hofmann-Engl 2010), and provides a clear overview of existing techniques. The analysis of the fifteen systems considered has allowed us to identify the strengths and weaknesses of the algorithms as well as wider trends in the field.

Starting from this point, we have been able to recommend avenues of future research that will hopefully lead to further improvement in the area. In this regard, we highlight the lack of a widely approved definition of melodic similarity, which strongly affects both the development of algorithms and the outcomes of competitions. We

recommend the development of a common ground-truth that can be used as a standard corpus for comparing systems. Also, we observed a worryingly small number of systems which are able to analyse polyphonic pieces. The capacity to analyse such music is of primary importance in fostering the exploitation of melodic similarity tools in real-world applications. Finally, because musical works belonging to distinct styles are often very significantly different, we have found that it is extremely difficult to develop systems that perform well across a range of styles. Therefore, we believe that future algorithms should have a relatively limited scope, and should be configured on a specific style of music, through machine-learning techniques.

References

Aloupis, G., T. Fevens, S. Langerman, T. Matsui, A. Mesa, Y. Nuñez, D. Rappaport, and G. Toussaint. 2006. "Algorithms for computing geometric measures of melodic similarity." *Computer Music Journal* 30(3):67–76.

Bohak, C., and M. Marolt. 2009. "Calculating Similarity of Folk Song Variants with Melody-based Features." In *proceedings of ISMIR*. pp. 597–602.

Cambouropoulos, E. 1998. "Towards a general computational theory of musical structure." Ph.D. thesis, University of Edinburgh.

Cleary, J. G., and I. Witten. 1984. "Data compression using adaptive coding and partial string matching." *Communications, IEEE Transactions on* 32(4):396–402.

de Carvalho Junior, A., and L. Batista. 2012. "Sms identification using PPM, psychophysiological concepts, and melodic and rhythmic elements." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Downie, J. S. 1999. "Evaluating a simple approach to music information retrieval: Conceiving melodic n-grams as text." Ph.D. thesis, The University of Western Ontario.

Downie, J. S. 2004. "The scientific evaluation of music information retrieval systems: Foundations and future." *Computer Music Journal* 28(2):12–23.

Downie, J. S. 2008. "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research." *Acoustical Science and Technology* 29(4):247–255.

Forte, A. 1973. *The structure of atonal music*. Yale University Press.

Forte, A., and S. E. Gilbert. 1982. *Introduction to Schenkerian Analysis: Form & Content in Tonal Music*. R.S. Means Company.

Frieler, K. 2006. "Generalized N-gram Measures for Melodic Similarity." In *Data Science and Classification*. pp. 289–298.

Frieler, K., and D. Müllensiefen. 2005. "The simile algorithm for melodic similarity." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Grachten, M., J. L. Arcos, R. Lopez de Mantaras, et al. 2004. "Melodic similarity: Looking for a good abstraction level." In *Proceedings of ISMIR*.

Hofmann-Engl, L. 2010. "An evaluation of melodic similarity models."

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning*. Springer.

Laitinen, M., and K. Lemström. 2010. "Geometric algorithms for melodic similarity." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Lemström, K. 2010. "Towards More Robust Geometric Content-Based Music Retrieval." In *proceedings of ISMIR*. pp. 577–582.

Lerdahl, F., and R. Jackendoff. 1985. *A generative theory of tonal music*.

Maidín, D. O. 1998. "A geometrical algorithm for melodic difference." *Computing in musicology: a directory of research* (11):65–72.

Marsden, A. 2012. "Interrogating melodic similarity: a definitive phenomenon or the product of interpretation?" *Journal of New Music Research* 41(4):323–335.

McNab, R. J., L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham. 1996. "Towards the digital music library: Tune retrieval from acoustic input." In *Proceedings of the first ACM international conference on Digital libraries*. pp. 11–18.

Meek, C., and W. P. Birmingham. 2002. "Johnny Can't Sing: A Comprehensive Error Model for Sung Music Queries." In *Proceedings of ISMIR*.

Müllensiefen, D., and K. Frieler. 2004. "Optimizing Measures Of Melodic Similarity For The Exploration Of A Large Folk Song Database." In *Proceedings of ISMIR*.

Müllensiefen, D., and K. Frieler. 2006. "Evaluating different approaches to measuring the similarity of melodies." In *Data Science and Classification*. pp. 299–306.

Narmour, E. 1992. *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press.

Orio, N., and A. Rodà. 2009. "A Measure of Melodic Similarity based on a Graph Representation of the Music Structure." In *Proceedings of ISMIR*. pp. 543–548.

Rizo, D., and J. M. Inesta. 2010. "Trees and combined methods for monophonic music similarity evaluation." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Roig, C., L. J. Tardón, A. M. Barbancho, and I. Barbancho. 2013. "Submission to Mirex 2013 Symbolic Melodic Similarity." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Suyoto, I. S., and A. L. Uitdenbogerd. 2005. "Simple efficient n-gram indexing for effective melody retrieval." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Suyoto, I. S., and A. L. Uitdenbogerd. 2010. "Simple orthogonal pitch with ioi symbolic music matching." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Urbano, J. 2013. "A Geometric Model supported with Hybrid Sequence Alignment." *Proceedings of the Annual Music Information Retrieval Evaluation exchange* .

Vempala, N. N., and F. A. Russo. 2015. "An Empirically Derived Measure of Melodic Similarity." *Journal of New Music Research* .

Wolkowicz, J., and V. Kešelj. 2011. "Text Information Retrieval Approach to Music Information Retrieval." .

Yazawa, S., Y. Hasegawa, K. Kanamori, and M. Hamanaka. 2013. "Melodic Similarity based on Extension Implication-Realization Model." *MIREX Symbolic Melodic Similarity Results* .