Fuzzy vectorial-based similarity detection of Music Plagiarism

Roberto De Prisco
University of Salerno
Salerno, Italy
Email: robdep@unisa.it

Delfina Malandrino
University of Salerno
Salerno, Italy
Email: dmalandrino@unisa.it

Gianluca Zaccagnino
University of Salerno
Salerno, Italy
Email: zaccagnino.gianluca@gmail.com

Rocco Zaccagnino
University of Salerno
Salerno, Italy
Email: zaccagnino@dia.unisa.it

***Abstract*—Plagiarism, i.e., copying the work of others and trying to pass it off as one own, is a debated topic in different fields. In particular, in music field, the plagiarism is a controversial and debated phenomenon that has to do with the huge amount of money that music is able to generate. However, the existing mechanisms for plagiarism detection mainly apply superficial and brute-force string matching techniques. Such well-known metrics, widely used to discover similarities in text documents, cannot work well in discovering similarities in music compositions.**

Despite the wide-spread belief that few notes in common between two songs is enough to decide whether a plagiarism exists, the analysis of similarities is a very complex process.

In this work, we provide novel perspectives in the field of automatic music plagiarism detection, and specifically, we propose an approach based on a fuzzy vectorial-based similarity.

Given a suspicious melody, our approach envisions three steps: *(1)* its transformation in a vectorial representation, *(2)* retrieving of a list of similar melodies, *(3)* analysis and comparison with this subset of associated similar scores by using a fuzzy degree of similarity, that varies in a range between 0 for melodies that are fully musically different, and 1 for identical melodies. To assess the effectiveness of our system we performed tests on a large dataset of ascertained plagiarisms. Results show that it is able to reach an accuracy of 93%.

I. Introduction

Plagiarism, which is the act of copying the work of others and trying to pass it off as one own is seen as a moral offense and often also a legal offense Plagiarism has an ancient root, as the word itself is derived from the Latin words *plagiaries*, which means abductor, and plagiare, which means to steal. Plagiarism has become a major concern since the establishment of education assessment. Since we entered the Internet era, the fast, vast, and easy access of information has further escalated the problem of plagiarism. It exists in many different scenarios, and is often difficult to prove or solve.

Examples of plagiarism include the textual plagiarism (verbatim copying, paraphrasing, and so on) applied for example to literature or to computer programs, and the artistic plagiarism, where examples are plagiarism in movie and television works or the misappropriation of the authorship of (parts of) musical compositions. Plagiarism occurs when two works are "*substantially similar*", whereas the owner of one of the two works has copied or has been inspired by the work of another. Interpreting or measuring the concept of substantially similarity is actually an open issue.

In this paper we address the topic of plagiarism in *pop music*. According to the legislation, there are no general and accepted rules according to which a minimum number of notes, or beats, simultaneously present in two music compositions constitutes plagiarism. Thus, over the years has emerged the need of defining the general criteria to identify cases of plagiarism. Plagiarism could be more fuzzy than clear, more complex than trivial copy and paste. In fact, conversely to what happen for textual plagiarism where a similarity can be easily identified between two or more elements, in the music field it is needed to consider all the features of a composition, that is, pitch, duration, intensity and timbre, as well as, melodic and harmonic dimensions.

Even not emphasized in the law, *melody* usually is the deciding element in cases of plagiarism. In fact, often we talk about *melodic plagiarism* to indicate music plagiarism. It is a phenomenon very debated in the music context, particularly in the context of western pop music. The recognition of a music plagiarism is not a trivial task, because it depends by the ability of recognizing similarities between melodic fragments inside two melodies.

In this work we propose a new music plagiarism detection method based on a fuzzy vectorial-based similarity. Given a suspicious melody, our approach envisions three steps: *(1)* its transformation in a vectorial representation, *(2)* retrieving of a list of similar melodies using the Jaccard coefficients, *(3)* analysis and comparison with this subset of associated similar scores. This stage entails the computation of a fuzzy degree of similarity, that varies in a range between 0 for melodies that are fully musically different, and 1 for identical melodies. Two melodies are marked as similar if they gain a fuzzy similarity score above a certain threshold. In order to compute the degree of similarity we introduce a new metric of melodic similarity, named "*vectorial melodic distance*" measure. To assess the effectiveness of our system we performed tests on a large dataset of ascertained plagiarisms. Results show that it is able to reach an accuracy of 93%.

The rest of the paper is organized as follows. In Section II we describe some relevant related works. In Section III we describe a new metric used to measure the similarity between two melodies. In Sections IV we provide details about a new music plagiarism detection method based on a fuzzy vectorial-based similarity. In Section V we describe the experimental

978-1-5090-6034-4/17/$31.00 ©2017 IEEE

analysis and, finally, in Section VI we conclude with final remarks and future directions.

II. Related work

Several approaches were proposed for complex music problems: machine learning [1], evolutionary algorithms [2], [3], [4], fuzzy logic [5] and so on. About the music plagiarism detection problem, most of the works proposed explore the using of well known string-matching techniques. These algorithms use various approaches towards the perception of similarity, but they can be generalized to either a spatial account [6] or a feature-based account [7]. One of the simplest algorithm used is the standard *edit distance measure* [8]. It compares two strings and computes the minimally required number of changes necessary to convert either string into the other one. Many contributions that apply it to melodies as strings of symbols can be found in [9], [10]. Based on the edit distance there is a slightly more sophisticated measure, the *weighted edit distance* [8], which is weighted by the duration of the notes. In [8] Müllensiefen and Frieler described other measures, that is the Ukkonen n-gram pitch measure and the Sum Common measure; the former sums differences of the frequencies of all the n-grams not occurring in both strings by reflecting the notion of difference between the two strings to be compared, while the latter, in contrast to the Ukkonen measure, sums the frequency of all n-grams occurring in both strings. Algorithms such as the *Earth Mover's Distance* [11] belong to a rather new approach developed to grasp similarity. The main idea is that "similarity is a function of the complexity required to distort or transform the representation of one into the representation of the other". The application of the Earth movers distance to music is accomplished by a mapping of (usually) one set time and pitch onto a subset of a two-dimensional space, the notes duration is visualized by the weight of each point. The similarity of two melodies is calculated by the amount of work necessary to change the location and weight of the point set representing one melody in order to convert it into the other melody [12]. An other approach is based on the *Tversky*'s metric [7]. In this case the similarity is based on the perception of features that two objects share and the salience of these features. People tend to compare the less salient stimulus with a more salient one and by establishing this order our judgments of similarity are shaped alike. The approaches based on this concept of similarity also include a weighting scheme developed from a large corpus of pop melodies [8].

We remark that most of these works are only focused on the melody and on a string representation of it. In our approach we propose a bidimensional vectorial representation of the melody in which we also considered the rhythmic information.

Most of the Fuzzy-based approaches available in literature literature were focused on text plagiarism. In [7], a copy detection approach for web documents was developed using a fuzzy information retrieval (IR) model. Koberstein and Ng [13] developed a reliable tool using fuzzy IR for determining the degree of similarity between two web documents and clustering the collection based on words. In addition, Alzahrani and Salim [14] adapted the fuzzy IR model for use with Arabic scripts.

About Fuzzy-based approaches for music plagiarism, several of these works focused on audio similarity, music copyright protection, personalized music recommender. Specifically, in [15] the authors proposed a method for rejection using fuzzy similarity measure and devise a music copyright protection system with the combination of this rejection method and a HMM-based music identification system. In [16] a system that combines metadata pattern matching and analysis of the music tracks to reach a more accurate point in recommending a song that a person may want to listen to, was introduced. Similarities among the songs are calculated based on Mel Frequency Cepstral Coefficients (MFCC) value of the songs. Using the similarity values of MFCC and appearance on playlists, a Mamdani Fuzzy Inference System was defined in order to give a weighted matrix which represents numerical distances among songs. The lower the distance between two songs is, the more similar they are. Finally, in [17] a new fuzzy similarity measures was introduces. Such a approach is based on the usage of spectrum histograms and fluctuation patterns audio fragments representation. By comparing these representations, is possible to determine the similarity between the corresponding fragments.

III. Measurement of melodic similarity

Finding an empirical approach towards melodic similarity is of great importance. Improving the measurement of melodic similarity will above all also progress our understanding of the perception of similarity in melodies.

In this section we describe a new metric used to measure the similarity between two melodies. To define a similarity measure it is necessary to first define a representation of the melodies and then a metric for measuring the similarity.

A. *Vectorial Melody representation*

In this section we provide, briefly and in a simplified way, the basic music notions needed to understand the rest of the paper. Each music note is associated to a *pitch*, a *name* and an *octave position*. Octaves are usually numbered from 0 to 6. In each octave there are 12 notes, named C, C# (or Db), D, D# (or Eb), E, F, F# (or Gb), G, G# (or Ab), A, A# (or Bb) and B. Usually, to identify a music note we only need of its name and its octave. In our approach, we prefer to use the pitch value of each note. As an example, C5 has pitch value 60. The *interval* between two notes is basically the distance of the pitches. Each note has a duration: $\frac{4}{4}$ (whole note), $\frac{2}{4}$ (half note), $\frac{1}{4}$ (quarter note), $\frac{1}{8}$ (eighth note), $\frac{1}{16}$ (sixteenth note), $\frac{1}{32}$ (thirty-second note) and $\frac{1}{64}$ (sixty-fourth note). A *rest note* is a note with pitch -1. A *melody* is a sequence of notes.

In this work we propose a new representation, named *vectorial representation*. Such a representation is based on the *interval* representation described in [18]. Let $M = (n_1, \ldots, n_k)$ be a melody where each n_i is a music note with $1 \leq i \leq k$. The

idea is to see a sequence of notes as a sequence of intervals. Thus, we could represent M with a string S_M obtained by using the interval (label) $interval(n_i, n_{i+1}) = n_{i+1} - n_i$ between n_i and n_{i+1} and by separating each interval by using a marker and by specifying each interval with $+$ if it is a positive interval and $-$ otherwise. If $n_i = -1$ (n_i is a rest note) or $n_{i+1} = -1$ (n_{i+1} is a rest note) then we set $interval(n_i, n_{i+1}) = 0$. For the rest note we use the special symbol p. We say that S_M is the *interval representation of* M [18]. As an example, let us consider the first two measures of the song entitled "Will you be there" of Michael Jackson, shown in Fig. 1,. As we can see, the sequence of notes is C4 (pitch 48), D4 (pitch 50), C4 (pitch 48), D5 (pitch 50), E4 (pitch 52), F4 (pitch 53), E4 (pitch 52) and finally F4 (pitch 53). Thus $M = (48, 50, 48, 50, 52, 53, 52, 53)$ and S_M = "$+2 * -2 * +2 * +2 * +1 * -1 * +1$". See [18] for further details.

data column= note, pitch

ni = pitch C4, ni+1 = pitch D4
intervall = ni+1 - n1 = 50 - 48 = +2

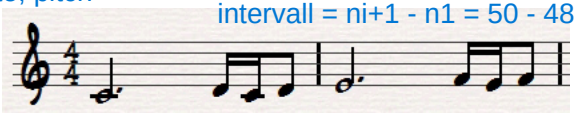

Fig. 1. First two measures of "Will you be there".

In our approach we extend the interval representation by defining a pair of vectors, named *vectorial representation of* M. Specifically, given the interval representation S_M of M, we define a vector V_M, named *melodic vector of* M, such that the value of each interval value (between to consecutive *) in S_M becomes a value in a V_M. Let us consider the previous example. We have $M = (48, 50, 48, 50, 52, 53, 52, 53)$ and S_M = "$+2 * -2 * +2 * +2 * +1 * -1 * +1$". Then the melodic vector of M is $V_M = [2, -2, +2, +2, +1, -1, +1]$. Second, we consider the rhythm inside the melody, i. e., information about duration of each note of melody. We remark that such a information should not be confused with the rhythm of the music composition, but as additional information about the melody. Given a melody M, the melodic vector V_M of M, we define a vector R_M, named *m-rhythmic vector of* M,such that $R_M.size() = V_M.size()$ and $R_M[i]$ contains the duration of the interval in $V_M[i]$, for each $0 \leq i \leq V_M.size()-1$. Let us consider the previous example. We have $M = (48, 50, 48, 50, 52, 53, 52, 53)$, S_M = "$+2 * -2 * +2 * +2 * +1 * -1 * +1$" and $V_M = [2, -2, +2, +2, +1, -1, +1]$. Then m-rhythmic vector of M is $R_M = [\frac{3}{4}, \frac{1}{16}, \frac{1}{16}, \frac{1}{8}, \frac{3}{4}, \frac{1}{16}, \frac{1}{16}, \frac{1}{8}]$.

vector of Sm

We say that the pair (V_M, R_M) is the vectorial representation of M. So, given a melody, we first compute the interval representation proposed in [18], and then we compute the corresponding vectorial representation.

B. *Vectorial Melodic distance*

As explained in Section I, one of the main problems about music plagiarism is that there are no general and accepted rules according to which a minimum number of notes, or beats, simultaneously present in two music compositions constitutes plagiarism. Thus, over the years has emerged the need of defining the general criteria to identify cases of plagiarism.

By a careful study of cases of plagiarism, it emerged that cases can be divided into 2 categories: *(1) text-like plagiarism class*, which is the set of cases in which melodies have many similar parts (can be detected with text plagiarism detection approaches); *(2) text-nolike plagiarism class*, which is the set of cases in which melodies have very few similar parts (cannot be detected with text plagiarism detection approaches). As an example of text-like plagiarism, we can consider "Love is a wonderful thing" of Micheal Bolton against "Love is a wonderful thing" of Isley Brothers. As we can see, in this case the melodies are very similar (also in the title!), and the court case ended with a sentence of plagiarism against Bolton. As an example of text-nolike plagiarism, we can consider "Will you be there" of Micheal Jackson against "I cigni di Balaka" of two famous Italian singers: Albano Carrisi & Romina Power. As we can see, in this case the melodies are very similar only in the first three-four measures, but also in this case the court case ended with a sentence of plagiarism against Jackson. Obviously, we need to define a new metric able to say that both cases in text-like plagiarism class and text-nolike plagiarism class are very similar.

In Fig. 2 we highlight the similar parts of the scores corresponding to such cases pf plagiarism. As we can see the cases are very different: in the first case scores are very similar (blue rectangles), otherwise in the second case scores contain few similar parts (red rectangles). Despite this, both the cases are sentenced as plagiarism.

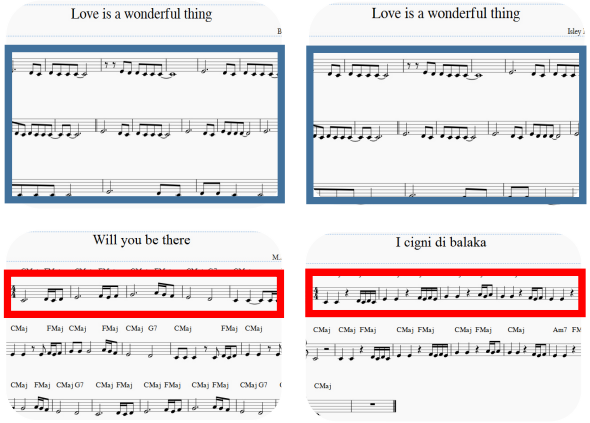

Fig. 2. Similar parts in the scores of plagiarism cases.

So the questions are: What is the rule? When a melody plagiarize another? How many similar parts there must be to have a plagiarism?

Thus, we need to introduce a definition of *distance* between vectorial representations of two melodies. Let M_1, M_2 be two melodies, let (V_{M_1}, R_{M_1}), (V_{M_2}, R_{M_2}) be the vectorial representations of M_1 and M_2, respectively. Let $\mathcal{S}_k(V_{M_1})$ and $\mathcal{S}_k(V_{M_2})$ be the sets of subvectors of V_{M_1} and of V_{M_2} respectively having size k, we compare elements of $\mathcal{S}_k(V_{M_1})$

with elements of $\mathcal{S}_k(V_{M_1})$. This corresponds to compare fragments of M_1 with fragments of M_2, from a melodic point of view.

The choice about the values of k is crucial since it must to be significant in musical terms. From the comparison point of view, we need to compare at least each measure of M_1 with each measure of M_2. At the same time we can to compare at most all the measures of M_1 (as a single fragment) with all the measures of M_2 (as a single fragment). Thus, let denote with $\mathcal{N}_{M_1,M_2}$ the minimum number of notes contained in a measure of M_1 and M_2, and $\mathcal{M}_{M_1,M_2} = \min(V_{M_1}.size(), V_{M_2}.size())$, we set:

$$\mathcal{N}_{M_1,M_2} \leq k \leq \mathcal{M}_{M_1,M_2}$$

Now, let $s_1 = [n_1, \ldots, n_k] \in \mathcal{S}_k(V_{M_1})$ and $s_2 = [m_1, \ldots, m_k] \in \mathcal{S}_k(V_{M_2})$ we compute the *weighted melodic vectorial distance* between s_1 and s_2, i.e., $\mathcal{D}_m(s_1, s_2) = \sqrt{\sum_{i=1}^{k} |n_i \cdot V_{M_1}.getPos(n_i) - m_i \cdot V_{M_2}.getPos(m_i)|^2}$. As we can see $\mathcal{D}_m(s_1, s_2)$ is the vectorial distance between s_1 and s_2 in which each n_i (resp. m_i) is weighted by the position inside M_1 (resp. M_2). This choice depends on the observation that for each case of plagiarism in the text-nolike class, the melodies have similar fragments at beginning, while the few similarity at the ending often are not considered by the court sentence. This means that the initial part of the melodies is considered most important.

Observe that, let $\mathcal{S}_k(R_{M_1})$ and $\mathcal{S}_k(R_{M_2})$ the set of subvectors of R_{M_1} and of R_{M_2} respectively having size k, for each $s_1 \in \mathcal{S}_k(V_{M_1})$ (resp. $s_2 \in \mathcal{S}_k(V_{M_2})$) there exists a corresponding $z_1 \in \mathcal{S}_k(R_{M_1})$ (resp. $z_2 \in \mathcal{S}_k(R_{M_2})$). So, given $s_1 = [n_1, \ldots, n_k] \in \mathcal{S}_k(V_{M_1})$ and $s_2 = [m_1, \ldots, m_k] \in \mathcal{S}_k(V_{M_2})$ and the corresponding subvectors $z_1 = [d_1, \ldots, d_k] \in \mathcal{S}_k(R_{M_1})$ and $z_2 = [t_1, \ldots, t_k] \in \mathcal{S}_k(R_{M_2})$ we compute the *weighted m-rhythmic vectorial distance* between s_1 and s_2, i.e., $\mathcal{D}_r(s_1, s_2) = \sqrt{\sum_{i=1}^{k} |d_i \cdot R_{M_1}.getPos(d_i) - t_i \cdot R_{M_2}.getPos(t_i)|^2}$.

So, given $s_1 \in \mathcal{S}_k(V_{M_1})$ and $s_2 \in \mathcal{S}_k(V_{M_2})$ we say that the *vectorial melodic distance between* s_1 and s_2 is $\mathcal{D}(s_1, s_2) = (\mathcal{D}_m(s_1, s_2), \mathcal{D}_r(s_1, s_2))$.

Let M_1, M_2 be two melodies, we consider the multi-objective problem of individuate the subvectors $s_1 \in V_{M_1}$ and $s_2 \in V_{M_2}$ with less vectorial melodic distance, among each $\mathcal{N}_{M_1,M_2} \leq k \leq \mathcal{M}_{M_1,M_2}$. As a solution we consider the Pareto front and we choose the pair of subvectors with the lower vectorial melodic distance, and at equal melodic distance we choose the pair of subvectors with the lower vectorial r-melodic distance. Let (s_1, s_2) be such a pair, we indicate with $\mathcal{D}(M_1, M_2) = \mathcal{D}(s_1, s_2)$ the *vectorial melodic distance between* M_1 and M_2.

IV. A FUZZY VECTORIAL-BASED SIMILARITY APPROACH

In this section, we describe a new music plagiarism detection method based on a new definition of *fuzzy vectorial-based similarity*. The problem considered is: given a suspicious melody M_x and a large collection of melodies $\mathcal{C}_M$, find all melodies $M_y \in \mathcal{C}_M$ that are similar to M_x according to such fuzzy semantic-based similarity.

A. The fuzzy framework

Our operational fuzzy framework is shown in Fig. 3. The process starts with a collection of melodies $\mathcal{C}_M$ and a suspicious melody M_x. Then: *(1)* we firstly compute the vectorial representation of M_x (see Section III-A), *(2)* we retrieve a list of similar melodies in $\mathcal{C}_M$ for M_x using the Jaccard coefficients, *(3)* M_x is analyzed and compared with this subset of associated similar melodies. This stage entails the computation of fuzzy degree of similarity that ranges between 0 for completely musically different scores and 1 for exactly musically identical scores. Two scores are marked as similar (the plagiarism exists) if they gain a fuzzy similarity score above a certain threshold. In order to compute the degree of similarity we use the vectorial melodic distance (see Section III-B).

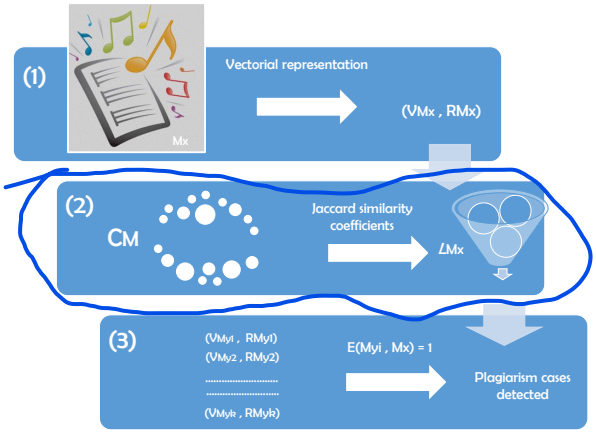

Fig. 3. Steps of the Fuzzy framework.

1) Step 1 - Preprocessing: In this step we compute the vectorial representation (V_{M_x}, R_{M_x}) of M_x as explained in Section III-A.

2) Step 2 - Retrieval of similar melodies: Near duplicate detection methods can be used to bring similar sources and discard dissimilar ones. We use shingling and Jaccard coefficient approach [19] applied to V_{M_x} and R_{M_x}. The k-gram referred to subvectors in $\mathcal{S}_k(V_{M_x})$ and $\mathcal{S}_k(R_{M_x})$. Intuitively, two melodies M_A and M_B are similar if they share enough k-grams. By performing union and intersection operations between the k-grams, we can find the melodic Jaccard similarity coefficient $\mathcal{J}(\mathcal{S}_k(V_{M_A}), \mathcal{S}_k(V_{M_B}))_m$ between V_{M_A} and V_{M_B} as stated in the following equation:

$$\mathcal{J}(\mathcal{S}_k(V_{M_A}), \mathcal{S}_k(V_{M_B}))_m = \frac{|\mathcal{S}_k(V_{M_A}) \cap \mathcal{S}_k(V_{M_B})|}{|\mathcal{S}_k(V_{M_A}) \cup \mathcal{S}_k(V_{M_B})|}$$

the m-rhytmic Jaccard similarity coefficient $\mathcal{J}(\mathcal{S}_k(R_{M_A}), \mathcal{S}_k(R_{M_B}))_r$ between R_{M_A} and R_{M_B} as stated in the following equation:

$$\mathcal{J}(\mathcal{S}_k(R_{M_A}),\mathcal{S}_k(R_{M_B}))_r = \frac{|\mathcal{S}_k(R_{M_A}) \cap \mathcal{S}_k(R_{M_B})|}{|\mathcal{S}_k(R_{M_A}) \cup \mathcal{S}_k(R_{M_B})|}$$

Therefore, let M_x be the suspicious melody, for each melody $M_y \in \mathcal{C}_M$ and for each $\mathcal{N}_{M_x,M_y} \leq k \leq \mathcal{M}_{M_1,M_2}$, we compute $\mathcal{J}(\mathcal{S}_k(V_{M_x}),\mathcal{S}_k(V_{M_y}))_m$ and $\mathcal{J}(\mathcal{S}_k(R_{M_x}),\mathcal{S}_k(R_{M_y}))_r$. Finally we consider the average melodic Jaccard similarity coefficient between M_x and M_y as:

$$\mathcal{J}(M_x,M_y)_m = \frac{\sum_{\mathcal{N}_{M_x,M_y} \leq k \leq \mathcal{M}_{M_x,M_y}} \mathcal{J}(\mathcal{S}_k(V_{M_x}),\mathcal{S}_k(V_{M_y}))_m}{\mathcal{M}_{M_1,M_2} - \mathcal{N}_{M_x,M_y} + 1}$$

and the average m-rhythmic Jaccard similarity coefficient between M_x and M_y as:

$$\mathcal{J}(M_x,M_y)_r = \frac{\sum_{\mathcal{N}_{M_x,M_y} \leq k \leq \mathcal{M}_{M_x,M_y}} \mathcal{J}(\mathcal{S}_k(R_{M_x}),\mathcal{S}_k(R_{M_y}))_r}{\mathcal{M}_{M_1,M_2} - \mathcal{N}_{M_x,M_y} + 1}$$

At this point we need to set a threshold t_m for the melodic Jaccard coefficient and a threshold t_r for the m-rhythmic Jaccard coefficient, in order to decide when M_x and M_y are similar. Formally, if $\mathcal{J}(M_x,M_y)_m \geq t_m$ and $\mathcal{J}(M_x,M_y)_r \geq t_r$ we consider M_y similar to M_x. We denote with $\mathcal{L}_{M_x}$ the list of melodies of $\mathcal{C}_M$ similar to M_x.

As we will see in Section V, the results of experiments have been better by setting $t_m = 0.2$ and $t_r = 0.15$.

3) Step 3: Fuzzy deep analysis: At this step, a detailed analysis between M_x and each melody $M_y \in \mathcal{L}_{M_x}$ is performed. To obtain the degree of similarity between M_x and M_y, a *fragment-to-melody correlation factor* for each fragment m_y of M_y and the melody M_x is computed as follows. Let z subvector of V_{M_y} and s subvector of V_{M_x} such that $\mathcal{D}(M_x,M_y) = \mathcal{D}(s,z)$ is the vectorial melodic distance between M_x and M_y (see Section III-B). Let k such that $s \in \mathcal{S}_k(V_{M_x})$ and $z \in \mathcal{S}_k(V_{M_y})$.

Then we use such a value k as size for the fragments to which calculate the fragment-to-melody correlation factor. In fact, for each $s_y \in \mathcal{S}_k(V_{M_y})$, $s_x \in \mathcal{S}_k(V_{M_x})$ we consider $d(s_y,s_x)$ as the normalized value of $\mathcal{D}(s_y,s_x)$ in the range $[0,1]$. We say that $d(s_y,s_x)$ is the *fuzzy similarity* between s_y and s_x. Thus the fragment-to-melody correlation factor for s_y and the melody M_x is computed as:

$$\lambda(s_y,M_x) = 1 - \prod_{s_x \in \mathcal{S}_k(V_{M_x})} (1 - d(s_y,s_x))$$

At this point we compute the *fuzzy vectorial-based similarity between* M_x and M_y as a fuzzy number between 0 and 1 using the following equation:

$$\mathcal{F}(M_y,M_x) = \frac{\sum_{s_y \in \mathcal{S}_k(V_{M_y})} \lambda(s_y,M_x)}{V_{M_y}.size()}$$

Now, to judge two melodies as plagiarized, the minimum similarity score should be above a threshold value α. Formally:

$$\mathcal{E}(M_x,M_y) = \begin{cases} 1 & \text{if } \min(\mathcal{F}(M_x,M_y),\mathcal{F}(M_y,M_x)) \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

As we will see in Section V, by experimental analysis on a large dataset of plagiarism cases it was found that best results were obtained with $\alpha = 0.75$.

V. Experimental Analysis

In this section we report the results of tests that we carried out to assess the effectiveness of our approach. For our experiments we used the dataset of plagiarism cases in [20]. As explained before, we indicate with $\mathcal{C}_M$ such a set. $\mathcal{C}_M$ contains more than 160 cases of plagiarism from 1900 to 2016. Each case is a pair (C_i,D_i) where C_i contains information about the complaint (year, the country in which the complaint was made, the court that issued the judgment, the detail of the complaining work such as author, title, audio, score and so on) and D_i contains information about the defending (year, the country in which the defense was made, the court that issued the judgment, the detail of the defending work such as author, title, audio, score and so on). We implemented a Java parser to convert each input file (audio, midi, pdf, musicxml) in the vectorial representation described in Section III-A. So, for each case (C_i,D_i) we have the vectorial representation of the melody M_{C_i} of C_i and the vectorial representation of the melody M_{D_i} of D_i.

The objective of our experiments was to assess the effectiveness of our approach by comparing his detection ability with some of the most used metrics used for detecting music plagiarism.

A. Experiments

To perform our experiments we first found the best configuration of our fuzzy detection system and then we compared such a configuration with other metrics.

1) Configuration of the fuzzy detection system: As explained in Section IV, to use the system presented in this work we need to set some parameters: the threshold t_m of the melodic Jaccard similarity coefficient, the threshold t_r of the m-rhythmic Jaccard similarity coefficient and the similarity score threshold α.

In order to find the value for t_m, t_r and α we run the system on $\mathcal{C}_M$ for values in $\{0.1, 0.15, 0.2, 0.25, \ldots, 0.9, 0.95\}$. For each value of α we compute the percentage of cases of plagiarism detected. The best performance have been obtained with $t_m = 0.2$ and $t_r = 0.15$, $\alpha = 0.75$, and a percentage of cases of plagiarism detected equal to 93%. In Table I we report the better configurations.

2) Comparison with other metrics: We have compared the detection ability of our fuzzy system with some of the most used metrics used for detecting music plagiarism. Each metric was tested on the dataset $\mathcal{C}_M$ and we computed the percentage of cases of plagiarism detected. Specifically, the metrics used are: Edit distance, Ukkonen measure, Sum Common measure, Tf-Idf correlation, Tverskys feature-based similarity. See [8]

TABLE I
CONFIGURATION OF THE FUZZY DETECTION SYSTEM

t_m	t_r	α	%
0.2	0.15	0.75	93%
0.15	0.2	0.5	88%
0.25	0.15	0.35	82%
0.1	0.1	0.25	75%
0.2	0.2	0.75	69%

TABLE II
COMPARISON WITH OTHER METRICS

Metric	Performance
Fuzzy vectorial	93%
Tverskys feature-based	90%
Tf-idf	82%
Ukkonen	74%
Sum common	72%
Edit distance	51%

for details on the using of such metrics for music similarity detection approaches.

The experiments were performed in the following way: for each metric, for each case $(C_i, D_i) \in \mathcal{C}_M$ we used the metric chosen to check if C_i D_i are similar. In Table II we report the result of the comparison. As we can see our approach (fuzzy vectorial) provide the best results, with a percentage of 93%.

VI. CONCLUSION

Plagiarism has become a major concern since the establishment of education assessment. It exists in many different scenarios, and is often difficult to prove or solve. Examples of plagiarism include the textual plagiarism applied for example to literature or to computer programs, and the artistic plagiarism, where examples are plagiarism in movie and television works or the misappropriation of the authorship of (parts of) musical compositions.

In this paper we address the topic of plagiarism in *pop music*. According to the legislation, there are no general and accepted rules according to which a minimum number of notes, or beats, simultaneously present in two music compositions constitutes plagiarism. Even not emphasized in the law, *melody* usually is the deciding element in cases of plagiarism. In fact, *melodic plagiarism* is a phenomenon very debated in the music context, particularly in the context of western pop music. The recognition of a melodic plagiarism is not a trivial task, because it depends by the ability of recognizing similarities between melodic fragments inside two melodies.

In this work, we provide novel perspectives on music plagiarism detection, and specifically, we propose a music plagiarism detection method based on a fuzzy vectorial-based similarity approach. Our performance measures on a large corpus of ascertained plagiarisms indicate that about 93% of our detections are correct.

As future work we will study new representations that allow to consider other aspects of music in addition to the melody: harmony, rhythm of the composition, text and so on. In addition we will perform new experiments on larger dataset to assess the effectiveness of our approach in other types of plagiarism, such as the text plagiarism.

REFERENCES

[1] R. De Prisco, A. Eletto, A. Torre, and R. Zaccagnino, "A neural network for bass functional harmonization," in *Applications of Evolutionary Computation, EvoApplications 2010: EvoCOMNET, EvoENVIRONMENT, EvoFIN, EvoMUSART, and EvoTRANSLOG, Istanbul, Turkey, April 7-9, 2010, Proceedings, Part II*, 2010, pp. 351–360.

[2] G. Acampora, J. M. Cadenas, R. De Prisco, V. Loia, E. M. Ballester, and R. Zaccagnino, "A hybrid computational intelligence approach for automatic music composition," in *FUZZ-IEEE 2011, IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, 27-30 June, 2011, Proceedings*, 2011, pp. 202–209.

[3] D. P. Roberto, G. Zaccagnino, and R. Zaccagnino, "A multi-objective differential evolution algorithm for 4-voice compositions," in *2011 IEEE Symposium on Differential Evolution, SDE 2011, Paris, France, April 11-15, 2011*, 2011, pp. 65–72.

[4] R. De Prisco, G. Zaccagnino, and R. Zaccagnino, "A Genetic Algorithm for Dodecaphonic Compositions," in *Applications of Evolutionary Computation - EvoApplications 2011: EvoCOMNET, EvoFIN, EvoHOT, EvoMUSART, EvoSTIM, and EvoTRANSLOG, Torino, Italy, April 27-29, 2011, Proceedings, Part II*, 2011, pp. 244–253.

[5] A. E. Yilmaz and Z. Telatar, "Potential applications of fuzzy logic in music." in *FUZZ-IEEE*. IEEE, 2009, pp. 670–675.

[6] R. N. Shepard, "Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space," *Psychometrika*, vol. 22, no. 4, pp. 325–345, 1957.

[7] A. Tversky, "Features of similarity." *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[8] D. Mllensiefen, "Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments," *Computing in Musicology*, vol. 13, pp. 147–176, 2004.

[9] M. Mongeaul and D. Sankoff, "Comparison of musical sequences," *Computers and the Humanities*, vol. 24, no. 3, pp. 161–175, Jun. 1990.

[10] T. Crawford, C. S. Iliopoulos, and R. Raman, "String-Matching Techniques for Musical Similarity and Melodic Recognition," *Computing in Musicology*, vol. 11, pp. 71–100, 1998.

[11] U. Hahn, N. Chater, and L. B. Richardson, "Similarity as transformation," *Cognition*, vol. 87, no. 1, pp. 1–32, 2003.

[12] R. Typke, F. Wiering, and R. C. Veltkamp, "Transportation distances and human perception of melodic similarity," *Musicae Scientiae*, vol. 11, no. 1_suppl, pp. 153–181, 2007.

[13] J. Koberstein and Y.-K. Ng, "Using word clusters to detect similar web documents," in *Proceedings of the First International Conference on Knowledge Science, Engineering and Management*, ser. KSEM'06, 2006, pp. 215–228.

[14] S. M. Alzahrani and N. Salim, "On the use of fuzzy information retrieval for gauging similarity of arabic documents," in *Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the*. IEEE, 2009, pp. 539–544.

[15] K. Kim, M. Lim, and J. Kim, "Music copyright protection system using fuzzy similarity measure for music phoneme segmentation," in *FUZZ-IEEE 2009, IEEE International Conference on Fuzzy Systems, Jeju Island, Korea, 20-24 August 2009, Proceedings*, 2009, pp. 159–164.

[16] M. S. Rahman, M. S. Rahman, S. U. I. Chowdhury, A. Mahmood, and R. M. Rahman, "A personalized music recommender service based on fuzzy inference system," in *15th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2016, Okayama, Japan, June 26-29, 2016*, 2016, pp. 1–6.

[17] K. Bosteels and E. E. Kerre, *Fuzzy Audio Similarity Measures Based on Spectrum Histograms and Fluctuation Patterns*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 213–231.

[18] R. D. Prisco, N. Lettieri, D. Malandrino, D. Pirozzi, G. Zaccagnino, and R. Zaccagnino, "Visualization of Music Plagiarism: Analysis and Evaluation," in *IV 2016, Lisbon, Portugal, July 19-22*, pp. 177–182.

[19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[20] C. Cronin, "Columbia Law School & UCLA Law Copyright Infringement Project," update: 2016. [Online]. Available: http://mcir.usc.edu/