Transcribing Debussy's *Syrinx* dynamics through Linguistic Description: the *MUDELD* algorithm

María Ros¹, Miguel Molina-Solana^{1,*}, Miguel Delgado¹, Waldo Fajardo¹, Amparo Vila¹

Department of Computer Science and Artificial Intelligence, Universidad de Granada, Spain

Abstract

Advances in computational power have enabled the manipulation of music audio files and the emergence of the *Music Information Retrieval* field. One of the main research lines in this area is that of *Music Transcription*, which aims at transforming an audio file into musical notation. So far, most efforts have focused on accurately transcribing the pitch and durations of the notes, and thus neglecting other aspects in the music score. The present work explores a novel line of action in the context of automatic music transcription, focusing on the dynamics, and by means of *Linguistic Description*. The process described in this paper (called *MUDELD*: **M**USIC **D**YNAMICS **E**XTRACTION through **L**INGUISTIC **D**ESCRPTION) departs from the data series representing the audio file, and requires the segmentation of the piece in phrases, which is currently done by hand. Initial experiments have been performed on eight recordings of Debussy's *Syrinx* with promising results.

Keywords: Linguistic Description, Music Dynamics, Dynamics transcription, *MUDELD*, Debussy *Syrinx*

1. Introduction

When performing a musical piece, musicians do not play exactly what is written in the score; they deviate from the music sheet and enrich the performance with tempo, dynamics and tuning deviations, among other aspects. It is precisely by this shaping that the music does sound alive and expressive rather than dull and plain [1, 2]. However, those deviations are just vaguely indicated in the music sheet—and in many occasions, they are not even indicated at all—and therefore it is the performer's duty to employ them according to her own aesthetic criteria and experience. Several studies and prototypes have been developed trying to gather and model that performing knowledge with the aim of better understanding the insights and processes, and enabling machines to play music pieces in a human-like way [3].

Within this context, another interesting problem—and still an open one—is that of music transcription (i.e. converting an audio recording into some form of musical notation) [4]. The task is complicated because the exact audio transcription is hardly representable as a meaningful score and thus, some simplifications and educated decisions shall be made to fit the audio in an understandable and realistic score. Whilst most efforts in automatic music transcription have been focused on pitch (which is considered completely solved in the monophonic case, and very advanced in the polyphonic one) [5] and duration, little effort has been done on trying to recover the original dynamic indications (marking the aforementioned deviations in volume) in the score. In fact, to the best of our knowledge there are no works on this matter. Our proposal is therefore the first one to address such an issue: departing from audio recordings, recovering the dynamic indications that might have been in the original music sheet. Therefore, this paper means a first step towards the broader context of tackling the issue of transcribing performing indications, extending the current focus on pitch and duration.

*Corresponding author

Email addresses: marosiz@decsai.ugr.es (María Ros), miguelmolina@ugr.es (Miguel Molina-Solana), mdelgado@ugr.es (Miguel Delgado), aragorn@ugr.es (Waldo Fajardo), vila@decsai.ugr.es (Amparo Vila)

¹All authors contributed equally to this work

Whereas music represented in a computer is numerical by nature —not in vain, a piece of music can be represented as series of numerical data, each one accounting for different musical features (pitch, energy, duration, ...)— the classic written notation is much closer to the natural language, with its inherent vagueness but higher legibility. The numerical notation is very suitable for computers, but not for humans. Referring to music dynamics in particular, a common and accepted set of linguistic expressions exist among musicians, but those terms are not consistently used, neither do they have a unique meaning for all individuals. Those indications are textual and symbolic indications, which mark, in a relative way, the sound intensity a part of the piece is to be played with. This vagueness in the indications poses an additional difficulty on the task of transcribing music dynamics.

Taking those facts into account, the approach we will follow in this article to annotate the numeric representation of a music recording into readable notation is based on *Linguistic Summarization and Description* techniques. These techniques aim at creating abbreviated, concise and human-consistent descriptions of data sets [6]. Linguistic Description will thus allow us to analyse the data series extracted from music recordings and automatically describe the dynamics (i.e. relative changes in volume) of every phrase in the piece with proper and adequate indications. To achieve this objective, our algorithm will additionally use fuzzy logic to handle the vague, imprecise and uncertain information from the music performance, increasing the robustness of the description methods [7].

The rest of this paper is organized as follows: Section 2 offers a review and a critical discussion on works related with the present paper, on both the Linguistic Description and Music areas. Section 3 formalizes the problem at hand and the different elements involved. Section 4 describes *MUDELD*, our proposal for addressing the transcription of music dynamics by means of linguistic description techniques, whilst Section 5 presents the experiments we have devised to test its validity and soundness. Finally, the paper concludes with a summary and an outline of futures lines of research and open problems (Section 6).

2. Related works

This section reviews related works in both the areas of *Linguistic Description*, and *Music Performance and Transcription*, with the aim of presenting the reader to the context in which the present work is framed.

2.1. Linguistic summarization and description

Linguistic Summarization and Description of Data is an emerging research area that aims at extracting knowledge from data, by means of natural language sentences understandable by humans. Due to the growing amount of data generated nowadays by different sources, this area is obtaining more and more relevance since it provides a mechanism to make more accessible the knowledge comprised on the stored data.

In particular, a linguistic data summary (or description) can be defined as a concise, human-consistent description of a (numerical) data set [6], intended to be *general*, *brief* and *accurate* [8]. Yager [6] was indeed the first one proposing a structure to build significant data summaries: “A summary of the data set D shall consist of three items: a summarizer S , a quantity in agreement Q , a measure of validity or truth of the summary, T ”. This approach was later extended and redefined by many others; in particular, by Kacprzyk and Zadrozny [9] who proposed the use of a protoform as a general form for any linguistic summary. Those extensions, some way or another, were all aimed at summarizing the contents of a database via natural language using linguistically quantified propositions. Not in vain, databases usually store huge amounts of information that is not easy to process.

It is relevant to note that those linguistic summaries commonly use Fuzzy Logic theory to model linguistic variables and incorporate different forms of imprecision in a collection of natural language sentences [10]. Recent developments in the Fuzzy Logic field have therefore their impact on Linguistic Description of Data, which also take advantage of Data Mining techniques by means of extracting dependencies and discovering association rules [11], with the final aims of either 1) summarizing the data, 2) describing properties contained in the data, or 3) eliminating redundancy. For instance, Rasmussen and Yager [12] studied the use of Gradual Functional Dependencies and proposed a query language called *SummarySQL* to evaluate the extracted dependencies among pieces of data. In [7], authors proposed the use of hierarchical conceptual clustering process to build the summary at different levels of abstraction obtaining a hierarchical summary to describe the database. Additionally, multidimensional databases [13] and type-2 fuzzy logic [14] have also been used within this context.

Apart from linguistic description in databases, summarization of data series (and time series in particular) has emerged as a fruitful research area, due to the nature of data series: continuous and flexible data, difficult to process

[15, 16]. A time series can be defined as a sequence of data distributed over a time span, and they have traditionally been studied to extract shape patterns where the composed data have a common specific meaning together. Those studies have usually been done by means of statistical methods [17], although they don't always produce a complete natural description for users.

Alternatively to statistical studies, Batyrshin et al. [18] advocated for the extraction of rule-based descriptions of time series using linguistic shape descriptors. Chiang, Chow and Wang [19] proposed a novel method for mining time series capable of overcoming the problem of having points very close or even equal to each other. In this line, Kacprzyk, Wilbik and Zadrozny [17] described a procedure to summarize temporal trends, identified with straight line segments of a piecewise linear. They use a modification of the well-known algorithm to find the partial trends by Sklansky and Gonzalez [20]. In their study, those partial trends were characterized with a set of features (attributes) used to obtain the linguistic summary of the time series. Castillo-Ortega, Marín and Sánchez [8] also proposed a method to summarize based on the fuzzy hierarchical partition of the time dimension, while taking into consideration the importance of subjectivity in a summary. These authors reported being able to achieve general, brief and accurate summaries in different domains. Some other works [21, 22, 23] have precisely advocated for segmenting the time dimension and summarize the data values within each sequence, focusing therefore in identifying frequent local cues.

While the focus of most works on linguistic summarization has been on stock market analysis and Finance applications, these techniques have proved very powerful when applied in other areas such as Image Analysis [24], Energy consumption [25], Weather forecast [26], or Elderly Living support [27].

2.2. Music performance

When skilled musicians play a piece of music, they do not do it mechanically, with constant tempo and intensity, exactly as written in the printed music score. Rather, they speed up at some places, slow down at others and stress certain notes. The most important parameters available to a performer are timing (tempo variations), dynamics (loudness variations) and articulation [28]. Changes in tempo (*timing*) are non-linear warping of the regular grid of beats that defines time in a score. *Dynamics* are changes in the sound intensity of notes with respect to the others and to the general energy of the fragment in consideration. *Articulation* consists in varying the gap between contiguous notes by, for instance, making the first one shorter or overlapping it with the next.

The way these parameters evolve during the performance is only loosely specified in the printed score (which basically just specifies the pitch, onset positions and durations, and has vague indications for the other aspects); because of that, the performer has freedom to shape the timings, dynamics and articulations according with her aesthetic and stylistic considerations [29]. Widmer and Goebel [1] precisely defined *expressive music performance* as “the deliberate shaping of the music by the performer, in the moment of playing, by means of continuous variations of parameters such as timing, loudness or articulation”. Advances in computational power have enabled the management and analysis of audio recordings through computers, and therefore, they have propitiated the computational modeling of music performance by means of analyzing music recordings. In this regard, there is an extensive literature on the analysis and modeling of music performances, focusing on different instruments and using different methodologies, as well as several attempts at computational representing such knowledge. We refer the interested reader to either [1] (the first one and still relevant), [2] (our own study) or [3] (the most recent). Attempts to use that knowledge to allow computers to expressively perform music also exists [30, 31].

Working with audio poses an additional interesting problem: the accurate transcription of the score departing from the audio, which is considered by many as one of the most relevant goals in the field of music signal analysis [4]. Although much work has been done in this regard, mostly solving the monophonic case, there are still problems with polyphonic music and particular instruments. The work by Poliner [5], for instance, discusses and evaluates several approaches for transcribing the melody.

However, as said before, music is not only a series of pitches and durations. Other aspects such as timing, dynamics and articulation are relevant and they are (although generally vaguely) indicated in the score. Precisely, whilst transcription efforts have mainly focused on recovering the notes and measures, a particular issue that has received little attention so far is that of recovering the dynamic indications from the audio. There have certainly been efforts at estimating the intensities of notes in a score-informed scenario (see Ewert and Muller's work [29]), but to the best of our knowledge, ours is the first work attempting an accurate transcription of the piece and phrase dynamics, employing the same indications that are in a written music sheet.

By taking advantage of the possibilities of representing a music performance as data series, our goal is to study their evolving shape and extract patterns that correspond to the musical dynamic indications, providing an accurate transcription of them. To do so, Linguistic Description will help to understand the numerical information, and provide linguistic descriptions, much more alike to the indications that can be found in the music sheets, and being therefore of greater usefulness to the musicians.

Summarization and description techniques have been used in the context of Music, although almost exclusively to the task of automatically providing a representative summary or ‘key phrase’ of a piece of music; that is, to find the most suitable and representative except that account for the whole piece [32]. This summarization, due to its particularities, is generally done by taking into account the low-level signal information from the audio. Głaczynski and Łukasik [33], for instance, discussed various approaches to automatic music audio summarization, and proposed a summarizing algorithm aiming at improving clarity, conciseness, coherence and usefulness. Whilst Chai [34] studied several methods for automatic music segmentation and summarization from audio signals and inquired scientifically into the nature of human perception of music, Jimenez et al. [35] aimed at finding the chorus of songs, which they considered the most representative part. Both reported promising results with their algorithms.

3. Problem definition and formalization

Having presented the background in which this work is framed, we devote this section to formally describe the problem at hand. As previously stated, our particular goal is to annotate the dynamics (changes in sound intensity) of a music performance with a set of linguistic expressions in such a way that they resemble the ones used by the musicians in the music scores. To achieve it, we firstly need to formally describe some related concepts which are also depicted in Figure 1.

Definition 3.1 (Music Performance). *A music performance M is a time series in which each observation $m(t) = (m_1(t), m_2(t), \dots, m_n(t))$, with $t \in T$, is a vector of n components. Each one of those components corresponds to a musical feature (e.g. pitch, loudness, timing...). M_i is the one-dimensional series corresponding to the observations $m_i(t)$, $\forall t \in T$.*

Definition 3.2 (Musical phrase). *A musical phrase P^j is a segment of a music performance M . A music performance M is a continuous concatenation of q musical phrases, $M = P^1 \parallel P^2 \parallel \dots \parallel P^q$. Partitions of M into phrases are generally performed following musical considerations.*

As aforementioned, Linguistic description is a research field gaining relevance lately thanks to their capacities to represent huge amount of data in an understandable way to humans [6, 36]. A complete definition of this concept and its associated process was provided by Yager [6]. Building upon his definition, we define the concept of linguistic annotation of a musical phrase:

Definition 3.3 (Linguistic annotation of a phrase). *Let be i one of the components of M . A i -related linguistic annotation of a phrase P^j , L_i^j , is a linguistic description of the data points $m_i(t)$ within P_i^j (phrase P^j , component i). It takes the form of a tuple $\langle S, Q, V \rangle$ where:*

- S is a linguistic expression interpreted by a fuzzy set and defined in the domain of a particular musical feature i . Based on [37], a linguistic expression is each one of the values that a linguistic variable can take in this domain (e.g. for $i = \text{loudness}$, $S \in \{\text{piano}, \text{mezzo}, \text{forte}\}$).
- Q is a quantifier; e.g. $Q \in \{\text{almost all}, \text{most}, \text{few}, \dots\}$ (see [38, 39]).
- $V \in [0, 1]$ is a measure of the validity of the description. The closer V is to 1, the more valid the proposed description is (see [6]).

As aforementioned, our objective consists of obtaining a concise and direct linguistic annotation for each musical phrase. In that sense, we will assume that Q is always the same quantifier for all annotations, and it will be omitted in the linguistic annotation structure, obtaining a simpler one: “Phrase P^j is S , with degree V ” (e.g. referring to music dynamics, “Phrase P^j is ‘piano’ with degree 0.8”). The particular procedure to calculate S , Q and V for each phrase P^j is specified in Section 4.

Figure 1: Visual representation of a music performance M and its different elements.

Figure 2: Visual representation of the process of performing a Linguistic Annotation.

Definition 3.4 (Linguistic annotation of a music performance). A i -related linguistic annotation of a music performance M , $A_i(M)$, is the concatenation of the i -related linguistic annotations L_i^j corresponding to all phrases P^j of M .

At this point we can formally describe our goal as follows: to devise a procedure to automatically obtain a i -related linguistic annotation of a given musical performance M , which is provided in the form of audio recording. This process is depicted in Figure 2.

Finally, in order to measure the correctness of the obtained *linguistic annotation*, we define a measure of *disagreement*, Dis , to quantify how far the automatically obtained dynamic indications are from the original indications in the music score (which we consider as correct reference values). Dis is based on the distance between points in the crisp domains in which the linguistic expressions were defined, and cannot be considered a distance measure as it does not satisfy the triangle inequality.

Definition 3.5 (Disagreement measure between two linguistic terms). Let $U \subset \mathbb{R}$ be an arbitrary domain, and L_r and L_s be linguistic terms interpreted by fuzzy sets, $\mu_{L_r}, \mu_{L_s} : U \rightarrow [0, 1]$. We define the disagreement between two linguistic terms L_r and L_s , $Dis(L_r, L_s)$, as:

$$Dis(L_r, L_s) = \inf\{d(x, y) \mid x \in \text{core}(L_r), y \in \text{core}(L_s)\} \quad (1)$$

with $d(x, y)$ being the distance in U between points x and y , and $\text{core}(L) = \{x \in U : \mu_L(x) = 1\}$.

We can also define the disagreement measure between a point $x \in U \subset \mathbb{R}$ and a linguistic term L_s (interpreted as a fuzzy set), $\text{Dis}(x, L_s)$, as the disagreement between the fuzzy singleton of x and L_s :

Definition 3.6 (Disagreement measure between a value and a linguistic term). Let $U \subset \mathbb{R}$ be an arbitrary domain, and L_s be a linguistic term interpreted by a fuzzy set, $\mu_{L_s} : U \rightarrow [0, 1]$. Let x be a point in U . We define the disagreement between x and L_s as:

$$\text{Dis}(x, L_s) = \inf\{d(x, y) \mid x \in \tilde{x}, y \in \text{core}(L_s)\} \quad (2)$$

Finally, and building upon the previous definitions, we can define the disagreement measure between two i -related linguistic annotations as:

Definition 3.7 (Disagreement between two linguistic annotations). Let $A_i(M)$ and $B_i(M)$ be two i -related linguistic annotations of a music performance M divided into q phrases $P^1, \dots, P^q$. Hence, linguistic annotations for their phrases can be named as A_i^j and B_i^j , $j \in \{1, \dots, q\}$. Following Definition 3.3, each A_i^j is a tuple that we will note as $\langle A_S, A_Q, A_V \rangle$ (and similarly, $\langle B_S, B_Q, B_V \rangle$ for B_i^j).

$$\text{Dis}(A_i^j, B_i^j) = \inf\{d(x, y) \mid \mu_{A_S}(x) = A_V, \mu_{B_S}(x) = B_V\} \quad (3)$$

We define the Disagreement between $A_i(M)$ and $B_i(M)$, $\text{Dis}_T(A_i, B_i)$, as the sum of disagreements between each phrase's corresponding annotations (using (3)).

$$\text{Dis}_T(A_i, B_i) = \sum_{j=1 \dots q} \text{Dis}(A_i^j, B_i^j) \quad (4)$$

4. MUDELD: Linguistic Description for transcribing music dynamics

Although expressive music performance comprises several aspects (see section 2.2: mainly timing, dynamic and articulation), we have constrained our study to only dynamics (i.e. relative variations in the loudness of the notes). The reason for this is to test our proposal in a limited scenario in which the vocabulary of linguistic expressions is well-defined and widely accepted. If results are successful, the procedure can be extended to other dimensions.

This section describes *MUDELD* (MUSIC Dynamics Extraction through Linguistic Description), the procedure we have devised to, departing from audio recordings, transcribe their dynamic indications (in the same way they are written in the music score) through Linguistic Description techniques. In brief, *MUDELD* is divided in the following steps, which will be described in detail in the rest of the current section, and particularized for a specific musical piece in the next one:

1. Represent an audio recording as a data series.
2. Select the relevant series.
3. Segment the piece in musical phrases.
4. Perform the description process on each phrase.

4.1. Audio representation

In order to computationally work with the music recordings, it is necessary to transform them to a representation that is manageable by a computer, usually some kind of numerical data. In the particular case of music, they can be further organized as data series. Many tools and encoding schemes exist to accommodate this computational representation, but they can mostly be divided in two groups: symbolic formats and audio formats.

Symbolic formats represent the musical events as objects and instructions (i.e. they do not describe actual sounds, but the score and instructions on how to play it), being much closer to the idea of music notation. Among the most popular formats within this category are *MIDI* and *MusicXML* [40].

Audio formats, on the other hand, encode the music directly as a set of sound samples along time. Some of these formats also include a compression procedure to reduce the file size that, in some cases, may incur in information

losses. Popular audio formats are *wav* (uncompressed) and *mp3* (with lossy compression). It is possible to convert between different audio formats, incurring in information losses depending on which formats are used.

As commercial music recordings are usually represented in audio formats, *MUDELD* will expect an audio format as input. Obtaining a textual numerical representation of those audio formats into data series is straightforward with the appropriate decoders, which are widely available in most programming languages. Therefore, the result of this step is a music performance M represented as a multidimensional data series, with each dimension accounting for a particular feature (pitch, timing, energy, velocity...).

4.2. Series selection

In order to extract the dynamics annotations from the recordings, we will focus our attention on the energy (i.e. sound intensity) of every note². As already indicated, our objective consists on producing a series of linguistic annotations about the dynamics of the different phrases in a performance, identifying both the general sound intensity in a phrase and its gradual changes. Luckily, in the music context, the sound intensity spectrum can be divided in several intervals, mainly *pianissimo* (**pp**), *piano* (**p**), *mezzo-piano* (**mp**), *mezzo-forte* (**mf**), *forte* (**f**), *fortissimo* (**ff**); along with the terms *crescendo* and *decrescendo* to account for the gradual changes in sound intensity.

As a music performance M (coming from an audio recording) is a multidimensional data series along several dimensions (energy, pitch, frequency...), it is necessary to focus only on the energy series. That can be done by means of the appropriate audio analysis tools, which are able to extract that particular data from the audio recordings. The output of this step is the music performance M reduced to only the series related with sound intensity: M_{energy} . In our particular case, we have obtained the sound intensity series of each one of the recordings in the dataset by means of Sonic Visualizer [41] and the *Energy* plugin within the *BBC Vamp plugins*³, which calculates the root mean square of energy (which is closely proportional to our perception of loudness) and low energy ratio of a signal.

Because the dynamic indications are relative to a particular piece (i.e. a *piano* indication does not represent an exact intensity value, it is just a relative indication), it is not useful for our goals to have raw values for the energy magnitude. Therefore, it is much more adequate to normalize the data series representing the raw loudness (M_{energy} in our case), and make their values range from 0 to 1, so that the scale is controlled. This normalization is performed independently for each performance, with 0 representing silence (minimum value) and 1 being the maximum energy value in a particular recording. Performing this transformation poses additional advantages as it will allow to compare the dynamics of different pieces in a standard scale.

4.3. Piece Segmentation

The current version of our algorithm works by means of performing the linguistic description process on each of the different phrases P^i that compose a performance M . Those phrases are formally sequences within the data series of the whole performance. Therefore, it is necessary to know the boundaries of those phrases, segmenting M .

Several techniques can be used to perform that segmentation, ranging from the manual to the automatic ones, and taking into consideration different stylistic criteria. A nice feature of our algorithm is its independence on the particular segmentation provided. Musically speaking, several dynamic changes might occur within one phrase. However, as the current version of the algorithm only returns one dynamic indication per phrase, it is desirable to use phrases with a reduced number of dynamic changes (ideally one) in order to obtain more accurate results.

4.4. Linguistic Description process. Energy and trend study

To perform the annotation, we aim to extract two types of dynamic indications for each phrase in a recording:

- S1. relative loudness of each phrase (using the linguistic expressions *piano*, *forte* and *mezzo*)
- S2. the gradual changes of loudness within phrases (namely *crescendo* and *decrescendo*).

²It is relevant to note that the terms *energy*, *sound intensity* (density of energy) and *loudness* (perception of sound intensity) are commonly used indistinguishably although they are slightly different.

³These plugins, implementing a collection of audio feature extraction algorithms, were designed and built during the 'Making Musical Mood Metadata (M4)' project [42] by Baume and Raimond, at the *BBC Research and Development* group. They are available at <https://github.com/bbcrd/bbc-vamp-plugins>

Consequently, according to section 3 we aim at obtaining linguistic descriptions in the form: P^j is $S1$ (and $S2$), where:

- P^j represents the j^{th} phrase in the performance.
- $S1$ corresponds to the *sound intensity*, the overall sound intensity of phrase P^j 's notes,
- $S2$ is the trend of the phrase. This term only appears in the linguistic description if the trend is stable.

Next, we define the set of linguistic variables required to perform the linguistic description process. Formally, Zadeh [37] defined a linguistic variable as a quadruple (X, T, U, M) , where X , represents the name of the linguistic variable, T , is the set of linguistic values (terms) that X can take, U , is the actual physical domain in which the linguistic variable X takes its quantitative (crisp) values and M , represents the semantic rule that relates each linguistic value in T with a fuzzy set in U .

As already indicated, musicians use many linguistic expressions to cover the whole range of dynamics of a musical phrase (i.e. the overall sound intensity of a phrase's notes); however, for the sake of simplicity and representativity, we constrained such a variety to only three terms (which are also the most commonly used): *piano* (p), *mezzo* (m)⁴, *forte* (f). Using this information, we define the linguistic variable of *sound intensity* over the sound intensity domain $[0, 1]$, with three linguistic terms: *piano*, *mezzo* and *forte*. Figure 3 shows the instantiation of *sound intensity* fuzzy variable as trapezoidal fuzzy sets over the (normalized) sound intensity domain.

Figure 3: Fuzzy sets for the *piano*, *mezzo* and *forte* labels, defined over the (normalized) *sound intensity* domain $[0, 1]$ (horizontal axis). The vertical axis represents the membership degree.

Secondly, we define a linguistic variable for the *gradual changes in volume* over the angle (in degrees) domain $[-90, 90]$. We identify three linguistic terms, corresponding to the different inclinations of a trend line: *crescendo*, *constant* and *decrescendo*, accounting for positive, flat and negative slopes, respectively. Figure 4 shows the instantiation of the linguistic variable *gradual changes in volume* as trapezoidal fuzzy sets.

Finally, we are interested in reporting the *gradual changes* only if they are stable within phrases. For measuring that stability we define another linguistic variable *Variability* over the mean square error domain $[0, 1]$ of the values with the slope, with the linguistic term *stable*. Figure 5 shows the fuzzy variable *Variability* defined as trapezoidal fuzzy sets for our particular problem.

Once we have determined the fuzzy variables and their fuzzy sets (which have been defined as a generalization of a wide range of cases studied, using expert knowledge), we have to define the method followed to extract the linguistic annotations. The process departs from the segmented numerical data series that represents the evolving sound intensity of a music performance in a particular audio recording. For each phrase P^j , which is processed independently, we calculate the three aforementioned values related with the *sound intensity*:

Average value: The average *sound intensity* value of the notes within a phrase is used to estimate a representative value for the set, although without taking into account the variability.

⁴Musicians do not usually use this level, but we will employ it to represent both *mezzo-piano* and *mezzo-forte*, which are the intermediate levels between *piano* and *forte*

Figure 4: Fuzzy sets for *crescendo* and *decrecendo* labels, described as the angle $[-90, 90]$ of the *Gradual Changes of volume* (horizontal axis). An additional label (*constant*) accounting for flat slopes has also been introduced. The vertical axis represents the membership degree.

Figure 5: Fuzzy sets for *Variability*. The horizontal axis indicates the MSE degree, in the interval $[0, 1]$. The vertical axis represents the membership degree.

Slope of the estimated linear equation: It is used to compute the angle of the *change in the dynamics* of each performance, providing the rate of change on the *sound intensity* values.

Mean square error (MSE): It determines the stability of each phrase, indicating if the calculated slope is in fact representative.

To calculate the slope of every phrase, we extrapolate here the concept of *dynamics of change* identified by Kaprzyk et al. [17], which represent the dynamics of changes by measuring the slope of a line representing the trend, quantifying the dynamics within the interval $[-90^\circ, 90^\circ]$ of possible angles. Based on our previous approach [43], we use the well-known method of linear regression to analyse every segment independently, with the aim of defining the trend within each one. We compute a predicted model (linear equation) from the observed data (musician's performance). Notice that, although we have preferred to use Kaprzyk et al. [17] proposal, there exists other alternatives to measure the slope of a line representing the trend (such as the one by Novak et al. [44] which extracts time series trends through the F-transform).

Due to the typology of datasets used, and in order to avoid false labels, we need to determine if there is an actual gradual change or just a series of sudden changes. We use the MSE measure calculated over the slope of the estimated linear approach for this purpose. If the obtained MSE can be considered stable, the slope will be representative and therefore, it can be given as part of the linguistic description. Otherwise, the slope indication should not be considered accurate, and thus will not be part of the output.

To sum up, each one of the defined linguistic variables are instantiated using the values calculated previously. In particular,

- *Sound intensity fuzzy variable* will be instantiated with the *Average value*
- *Gradual changes in volume* is compute as the slope of the estimated linear equation.
- *Variability* is calculated by the mean square error (MSE)

#	Performer	ASIN
(a)	Alain Marion	B000QN95OS
(b)	Alison Balsom	B000HKD7OY
(c)	André Noiret	B002ZO8ZZQ
(d)	Philippa Davies	B000TE6ECA
(e)	Emmanuel Pahud	B000THG1QQ
(f)	Wellington Cardoso	B000SH1UPE
(g)	James Galway	B001BKRN9Y
(h)	Marina Piccinini	B000QLDP8C

Table 1: Performer and ASIN’s unique reference of each one of the eight performances used in this study

5. Experimentation

In order to test the validity of the procedure described in the former section, we have tested *MUDELD* with a real music piece. This section describes the particular experimentation we have done, providing actual values, measures and results. The piece we used for the experimentation is Debussy’s *Syrinx* for solo flute (L. 129). This piece (see Appendix A for its score), an indispensable part of any flautist’s repertoire, gives the performer plenty of room for interpretation and emotion, and has been recorded many times by several players. Because of that, it is a well-known piece within the academia, with several works dealing with it [45, 46, 47, 48, 49]. *Syrinx* has rather complex dynamics, including frequent variations in sound intensity, being thus a challenging scenario for our proposal. Although the score includes plenty of annotations, we will only focus on the dynamics ones and, particularly, in the expressions we have indicated in Section 4.4.

When analysing audio recordings, two approaches are available. The first one employs material specifically recorded for the intended study and related to the performance aspect to be analyzed, emphasizing specific expressive resources. Alternatively, commercial recordings can be used, avoiding any bias towards the particular study and allowing a wider range of materials. As drawbacks, the scenario is not controlled and the sound analysis may be more complicated. Having that in mind, we compiled a dataset⁵ comprising eight different performances of *Syrinx* (see Table 1 for further details). These recordings were selected randomly from those in the search result of *Spotify*⁶. Henceforth, they came from different performers and recording conditions, and we have not interfered at all with the process of performing.

As previously stated, it is necessary to perform a segmentation of the piece into phrases and apply the linguistic description procedure to each one of those. Because the automatic segmentation of an arbitrary music piece is not yet completely solved, the experimentation presented in this paper rely on a manual segmentation in phrases (which can be found in Appendix A) performed by an expert musician. As stated in the previous section, this segmentation was done following musical criteria, but also aiming to minimize the number of dynamic changes within each phrase (which was not always possible). Other segmentations of *Syrinx*, mostly motivic ones, have been reported in literature (see [45, 47, 48, 49]).

Figure 6 shows the temporal evolution of the (normalized) energy values for the eight performances in the study, along with their segmentation in phrases. As can be seen, the performances have different temporal durations, and it is not immediately easy for a human expert to indicate the associate dynamics.

5.1. Results and discussion

Applying the *MUDELD* algorithm we obtained, for each *Syrinx* performance, a list of 19 dynamic annotations (one per phrase). In order to analyze the results for the complete experiment, Table 2 summarizes those results indicating which labels were selected for each phrase, taking into account the eight recordings in the study. For instance, the first phrase of *Syrinx* was identified as *piano* in one of the performances, *mezzo* in three, and *forte* in four; phrase 14, was in all cases classified as *forte*.

⁵To get access to it, we encourage the reader to contact the authors

⁶<https://www.spotify.com>

Figure 6: Representation of temporal evolution of the energy over time for each one of the eight (*a* to *h*) performances selected for the study. The vertical green lines represent the time points in which a new phrase begin. The energy (vertical) axis is normalized, whilst the time (horizontal) is not.

	phrase#																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
<i>forte</i>	4	3	0	1	0	1	3	0	0	0	2	2	4	8	7	0	0	0	0
<i>mezzo</i>	3	4	4	7	4	7	5	3	3	3	3	4	4	0	1	3	1	1	1
<i>piano</i>	1	1	4	0	4	0	0	5	5	5	3	2	0	0	0	5	7	7	7

Table 2: Summary of identified dynamic labels for each phrase, taking into account the population of eight performances available. Each cell in the table indicates the number of times that label has been selected. All columns sum eight. Each phrase’s value in gray background indicates the dynamic label from the original score, whereas bold text indicates the most voted labels in each phrase. The only phrase in which the correct label does not coincide with the mode (the most repeated label) is *phrase 1*

The first conclusion we get from Table 2 is that in all cases, the most voted linguistic annotation for each phrase is always but the first phrase, the one that arises from the score. Also, we can see that there are four phrases (3, 5, 11, 13) in which there are two labels with the same number of occurrences. There are two possible explanations for this fact: 1) a deeper segmentation was probably necessary, or 2) that phrase is ambiguous in its dynamics and can be played in several ways.

Next, we wanted to evaluate the quality of the dynamic-related annotations obtained from the *MUDELD* algorithm. To do that, we computed how closed were they to the dynamic indications in the original music score (which we consider the correct reference values). Using Equation (1), we calculate the disagreement between two linguistic annotations for the same phrase. Extending this concept, we calculate, through Algorithm 1 and following Definition 3.7, the disagreement between two linguistic annotations of a music performance.

Algorithm 1 Disagreement measure between linguistic annotations of a performance

Definitions:

$P^1, \dots, P^q$: phrases of a music performance M

L : set of linguistic labels defined over $U \subset \mathbb{R}$

$Dis(x, L_s)$: Disagreement between a point $x \in U$ and a label L_s (see Definition 3.6)

$maxDis_s$: $\max\{Dis(y, L_s) \mid y \in U, L_s \in L\}$

Inputs:

$A(M)$: linguistic annotation of performance M (*MUDELD*’ labels)

$R(M)$: linguistic annotation of performance M (reference labels)

$x^j \in U$: the crisp value associated with A^j

Outputs:

Dis_T : disagreement between annotations $A(M)$ and $R(M)$

Algorithm:

for all phrases $P^j, j = 1 \dots q$ **do**

$d^j = Dis(x^j, R^j) / maxDis_{R^j}$

end for

$Dis_T = \sum d^j / q$

To reinforce the results obtained, we computed Cohen’s κ coefficient [50], which calculate a degree of agreement between the annotations by *MUDELD* and the ones from the score. This coefficient also provided a method to evaluate our measure, by comparing it with this one. Cohen’s κ , ranging between -1 to 1, offers a conservative estimation of the agreement between two judges classifying N items (19 phrases) in C mutually exclusive qualitative classes (3 dynamic labels). Landis and Koch [51] proposed, not without controversy, a qualitative scale of significance to κ : values below 0 means no agreement, values within $[0.21 - 0.4]$ are considered as fair agreement, within $[0.41 - 0.6]$ as moderate, $[0.61 - 0.8]$ as substantial, and $[0.81 - 1]$ as almost perfect agreement. Table 3 summarizes the partial and total Dis , and the Cohen’s κ coefficient among each recording in the dataset and the reference annotation.

	phrase#																			Dis_T	κ
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19		
(a)	0	0.19	0.44	0.64	0	0	0.39	0.39	0.38	0.43	0.33	0.22	0	0	0	0.44	0.36	0.34	0.27	4.8	0.01
(b)	0	0	0	0.27	0	0	0.18	0	0	0	0	0	0.10	0	0	0	0	0	0	0.5	0.83
(c)	0.41	0.23	0.15	0	0	0.20	0	0.11	0.13	0.39	0.42	0.46	0	0	0	0.34	0	0	0	2.8	0.24
(d)	0.21	0	0	0	0.15	0	0	0	0	0	0.38	0.35	0.11	0	0	0	0	0	0	1.2	0.66
(e)	0.37	0.30	0	0	0.34	0	0	0	0	0	0.16	0	0	0	0	0	0	0	0	1.2	0.66
(f)	0.51	0	0.10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.6	0.83
(g)	0.72	0.50	0.26	0	0	0	0.19	0.15	0.24	0.24	0	0	0.11	0	0	0.46	0	0	0	2.9	0.35
(h)	0	0	0	0	0.21	0	0	0	0	0	0.32	0.16	0.09	0	0.37	0	0	0	0	1.2	0.73
Σ	2.22	1.21	0.95	0.64	0.97	0.2	0.75	0.65	0.75	1.05	1.62	1.20	0.41	0	0.37	1.23	0.36	0.34	0.27		
$mean$	0.28	0.15	0.12	0.08	0.12	0.03	0.09	0.08	0.09	0.13	0.2	0.15	0.05	0	0.05	0.15	0.05	0.04	0.03		

Table 3: Disagreement (Dis_T column), in the range [0-1], between each performance (a to h) and the score (reference values). The partial disagreement for each phrase are also shown (columns 1 to 19). Dis is based on (1) and calculated following Algorithm 1. Column κ shows Cohen's κ appa values with the agreement between the outputs of *MUDEL*D and the original score. Last two rows indicate respectively the sum and the mean of values in the corresponding columns; a zero value means that the *MUDEL*D annotations for that phrase match the one of the original score.

Analysing Table 3, we can conclude that our proposal is robust and valid, since it obtains quite similar results matching the dynamic indications on the written score. We can see how the percentage of disagreement for each phrase does never exceed 30%, obtaining in most cases values below 10%. Moreover, Table 3 shows that both *Dis* and κ are in agreement, validating the claim that our devised disagreement measure is a valid indicator of the agreement between the dynamic indications of two recordings. According to both *Dis* and κ columns, we can conclude that performances (*b*) and (*f*) are the most loyal to the written score; on the other hand, performance (*a*) is the most dissimilar. Further studies, from a musicological point of view, could be done in order to understand the reasons (if any) of those differences (e.g. source score, period, school, ...).

Regarding the gradual changes in sound intensity, the obtained results were not satisfactory. As we enforced a minimum degree of stability in the series to guarantee that the *crescendo* and *decrescendo* were truly there, it turned out that the current segmentation provoked that only a few of the phrases were detected as having a *crescendo* or *decrescendo*. Only the very clear cases covering the whole phrase (e.g. phrase 7 having a *decrescendo*) were detected.

6. Conclusions and future work

This paper has presented a novel application of Linguistic Description techniques towards the automatic transcription of music dynamics from audio recordings. After presenting the problem and related works, we have defined the linguistic variables and several fuzzy sets to account for the music dynamics, and have developed a procedure (*MUDELD*), based on Linguistic Description techniques to transform the raw numerical data representing the music recordings into those labels.

We have applied *MUDELD* to eight different recordings of Debussy's *Syrinx* (a rather complex scenario) in order to demonstrate the potential of this proposal to transcribe the music dynamics of this piece. From the performed experiments we can conclude that the proposal is certainly valid and sound. The most voted annotation is always (but in one phrase out of 19) matching the dynamic indications on the written score, and only in four of those phrases there is another annotation equally likely. A disagreement measure between dynamics of performances has been also proposed and applied to the dataset at hand, comparing how similar the outputs from *MUDELD* were from the original score indications. We found that seven of the eight studied performances were very loyal to the written indications in the score.

However, we acknowledge that, as a first attempt at transcribing the dynamics of a piece of music, the current version of *MUDELD* still needs improvements and further research. In the next future, we aim at automating the whole process of transcribing the dynamics of musical pieces. To achieve such a goal, an automatic segmentation of the piece in phrases is required, in contrast with our current study in which that task is manually made. Being aware of the current state-of-the-art on automatic music segmentation (with algorithms not being fully accurate yet), we however believe that their outputs might be good enough for the Linguistic Description process to obtain satisfactory annotations.

Because the segmentation poses concerns in itself (for instance, why should the dynamics annotations be constrained to only one in each phrase), we expect to use an adaptive sliding window in the Linguistic Description process to detect sudden changes in dynamics. This way, a phrase segmentation is no longer needed as the algorithm will be able to provide annotations corresponding to the passages in which dynamics maintain a tendency. Not in vain, if the dynamics in the audio change, there is likely a written indication at that point in the music score. Besides those modifications in the process, we also intend to apply and test *MUDELD* on a different set of musical pieces to check its validity on other pieces and authors. Because, the procedure does not rely on any piece-specific aspects, we anticipate that the results will very likely be consistent with the current ones. All in all, we expect the results and methodologies in this work to be inspiring for other researchers interested on the issue of music dynamic transcription.

Acknowledgements

This work is partly funded by the Spanish Government under project TIN2012-30939.

Appendix A. *Syrinx*'s score with phrase segmentation and dynamic labels

Syrinx

à Louis Fleury

Claude Debussy (1913)

Flûte seule

1 *Très modéré*

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

The musical score for *Syrinx* for flute solo is presented in a single system with 65 measures. The key signature is B-flat major (two flats), and the time signature is 3/4. The score is divided into eight numbered phrases, each marked with a red vertical line. The tempo is marked 'Très modéré' at the beginning. Dynamic markings include *mf* (mezzo-forte), *m* (mezzo), *p* (piano), and *Retenu* (sustained). The score includes various musical notations such as slurs, ties, and triplets. The first phrase (measures 1-4) starts with *mf* and ends with a red line. The second phrase (measures 5-8) starts with *mf* and ends with a red line. The third phrase (measures 9-12) starts with *mf* and ends with a red line. The fourth phrase (measures 13-16) starts with *p* and ends with a red line. The fifth phrase (measures 17-20) starts with *Retenu* and ends with a red line. The sixth phrase (measures 21-24) starts with *m* and ends with a red line. The seventh phrase (measures 25-28) starts with *m* and ends with a red line. The eighth phrase (measures 29-32) starts with *p* and ends with a red line.

[illegible]

References

References

- [1] G. Widmer, W. Goebl, Computational models of Expressive Music Performance: the state of the art, *Journal of New Music Research* 33 (3) (2004) 203–216. doi:10.1080/0929821042000317804.
- [2] M. Delgado, W. Fajardo, M. Molina-Solana, A state of the art on computational music performance, *Expert Systems with Applications* 38 (1) (2011) 155–160. doi:10.1016/j.eswa.2010.06.033.
- [3] A. Kirke, E. R. Miranda (Eds.), *Guide to Computing for Expressive Music Performance*, Springer, 2013. doi:10.1007/978-1-4471-4123-5.
- [4] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, A. Klapuri, Automatic music transcription: challenges and future directions, *Journal of Intelligent Information Systems* 41 (3) (2013) 407–434. doi:10.1007/s10844-013-0258-3.
- [5] G. E. Poliner, D. P. Ellis, A. F. Ehmann, E. Gomez, S. Streich, B. Ong, Melody transcription from music audio: Approaches and evaluation, *IEEE Transactions on Audio, Speech, and Language Processing* 15 (4) (2007) 1247–1256. doi:10.1109/TASL.2006.889797.
- [6] R. R. Yager, A new approach to the summarization of data, *Information Sciences* 28 (1) (1982) 69–86. doi:10.1016/0020-0255(82)90033-0.
- [7] G. Raschia, N. Mouaddib, SAINTETIQ: A fuzzy set-based approach to database summarization, *Fuzzy Sets and Systems* 129 (2) (2002) 137–162.
- [8] R. Castillo-Ortega, N. Marín, D. Sánchez, A fuzzy approach to the linguistic summarization of time series, *Multiple-Valued Logic and Soft Computing* 17 (2–3) (2011) 157–182.
- [9] J. Kacprzyk, S. Zadrozny, Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools, *Information Sciences* 173 (4) (2005) 281–304. doi:10.1016/j.ins.2005.03.002.
- [10] J. Kacprzyk, S. Zadrozny, Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation, *IEEE Transactions on Fuzzy Systems* 18 (3) (2010) 461–472. doi:10.1109/TFUZZ.2010.2040480.
- [11] J. Cubero, J. Medina, O. Pons, M. Vila, Data summarization in relational databases through fuzzy dependencies, *Information Sciences* 121 (34) (1999) 233–270. doi:10.1016/S0020-0255(99)00104-8.
- [12] D. Rasmussen, R. R. Yager, Finding fuzzy and gradual functional dependencies with SummarySQL, *Fuzzy Sets and Systems* 106 (2) (1999) 131–142. doi:10.1016/S0165-0114(97)00268-6.
- [13] A. Laurent, A new approach for the generation of fuzzy summaries based on fuzzy multidimensional databases, *Intelligent Data Analysis* 7 (2) (2003) 155–177.
- [14] A. Niewiadomski, A type-2 fuzzy approach to linguistic summarization of data, *IEEE Transactions on Fuzzy Systems* 16 (1) (2008) 198–212.
- [15] M. Delgado, W. Fajardo, M. Molina-Solana, Representation model and learning algorithm for uncertain and imprecise multivariate behaviors, based on correlated trends, *Applied Soft Computing* doi:10.1016/j.asoc.2015.07.033.
- [16] G. E. Box, G. M. Jenkins, G. C. Reinsel, *Time series analysis: forecasting and control*, John Wiley & Sons, 2013.
- [17] J. Kacprzyk, A. Wilbik, S. Zadrozny, Linguistic summarization of time series using a fuzzy quantifier driven aggregation, *Fuzzy Sets and Systems* 159 (12) (2008) 1485–1499, *advances in Intelligent Databases and Information Systems*. doi:10.1016/j.fss.2008.01.025.
- [18] I. Batyrshin, R. Herrera-Avelar, L. Sheremetov, A. Panova, Moving approximation transform and local trend associations in time series data bases, in: I. Batyrshin, J. Kacprzyk, L. Sheremetov, L. A. Zadeh (Eds.), *Perception-based Data Mining and Decision Making in Economics and Finance*, Vol. 36 of *Studies in Computational Intelligence*, Springer Berlin Heidelberg, 2007, pp. 55–83. doi:10.1007/978-3-540-36247-0_2.
- [19] D.-A. Chiang, L. Chow, Y.-F. Wang, Mining time series data by a fuzzy linguistic summary system, *Fuzzy Sets and Systems* 112 (3) (2000) 419–432.
- [20] J. Sklansky, V. Gonzalez, Fast polygonal approximation of digitized curves, *Pattern Recognition* 12 (5) (1980) 327–331. doi:10.1016/0031-3203(80)90031-X.
- [21] V. Novák, V. Pavliska, M. Štěpnička, L. Štěpničková, Time series trend extraction and its linguistic evaluation using F-Transform and fuzzy natural logic, in: L. A. Zadeh, A. M. Abbasov, R. R. Yager, S. N. Shahbazova, M. Z. Reformat (Eds.), *Recent Developments and New Directions in Soft Computing*, Vol. 317 of *Studies in Fuzziness and Soft Computing*, Springer International Publishing, 2014, pp. 429–442. doi:10.1007/978-3-319-06323-2_27.
- [22] D. Downing, V. Fedorov, W. Lawkins, M. Morris, G. Ostrochov, Large data series: modeling the usual to identify the unusual, *Computational Statistics & Data Analysis* 32 (3–4) (2000) 245–258. doi:10.1016/S0167-9473(99)00079-1.
- [23] B. Saleh, F. Massegli, Discovering frequent behaviors: time is an essential element of the context, *Knowledge and Information Systems* 28 (2) (2010) 311–331. doi:10.1007/s10115-010-0361-5.
- [24] A. Alvarez-Alvarez, D. Sanchez-Valdes, G. Trivino, Automatic linguistic description about relevant features of the mars’ surface, in: *International Conference on Intelligent Systems Design and Applications (ISDA)*, Córdoba, Spain, 2011, pp. 154–159.
- [25] A. Van Der Heide, G. Triviño, Automatically generated linguistic summaries of energy consumption data, in: *International Conference on Intelligent Systems Design and Applications (ISDA)*, Pisa, Italy, 2009, pp. 553–559.
- [26] A. Ramos-Soto, A. Bugarin, S. Barro, J. Taboada, Automatic generation of textual short-term weather forecasts on real prediction data, in: *LNC3, International Conference on Flexible Query Answering Systems 2013*, Vol. 8132, Springer Berlin Heidelberg, Granada, Spain, 2013, pp. 269–280. doi:10.1007/978-3-642-40769-7_24.
- [27] A. Wilbik, J. Keller, A fuzzy measure similarity between sets of linguistic summaries, *Fuzzy Systems, IEEE Transactions on* 21 (1) (2013) 183–189. doi:10.1109/TFUZZ.2012.2214225.
- [28] P. N. Juslin, J. A. Sloboda, *Music and emotion*, Oxford University Press, New York, 2001.
- [29] S. Ewert, M. Muller, Estimating note intensities in music recordings, in: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 385–388. doi:10.1109/ICASSP.2011.5946421.

- [30] R. López de Mántaras, Playing with cases: Rendering expressive music with case-based reasoning, *AI Magazine* 33 (4) (2012) 22–32. doi:10.1609/aimag.v33i4.2405.
- [31] S. Flossmann, M. Grachten, G. Widmer, Expressive Performance Rendering with Probabilistic Models, Springer, 2013, pp. 75–98.
- [32] B. Logan, S. Chu, Music summarization using key phrases, in: *Procs. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, IEEE, Istanbul, Turkey, 2000, pp. 749–752.
- [33] J. Głaczynski, E. Łukasik, Automatic Music Summarization. A “Thumbnail” Approach, *Archives of Acoustics* 36 (2) (2011) 297–309. doi:10.2478/v10168-011-0023-y.
- [34] W. Chai, Semantic segmentation and summarization of Music, *IEEE Signal Processing Magazine* 23 (2) (2006) 124–132. doi:10.1109/MSP.2006.1598088.
- [35] A. Jiménez, M. Molina-Solana, F. Berzal, W. Fajardo, Mining transposed motifs in music, *Journal of Intelligent Information Systems* 36 (1) (2011) 99–115. doi:10.1007/s10844-010-0122-7.
- [36] J. Kaprzyk, R. R. Yager, Linguistic summaries of data using fuzzy logic, *International Journal of General Systems* 30 (2) (2001) 133–154. doi:10.1080/03081070108960702.
- [37] L. Zadeh, The concept of a linguistic variable and its application to approximate reasoningii.
- [38] V. Novák, A formal theory of intermediate quantifiers, *Fuzzy Sets and Systems* 159 (10) (2008) 1229–1246. doi:10.1016/j.fss.2007.12.008.
- [39] M. Delgado, M. D. Ruiz, D. Sánchez, M. A. Vila, Fuzzy quantification: a state of the art, *Fuzzy Sets and Systems* 242 (2014) 1–30. doi:10.1016/j.fss.2013.10.012.
- [40] M. Good, MusicXML for notation and analysis, The virtual score: representation, retrieval, restoration 12 (2001) 113–124.
- [41] C. Cannam, C. Landone, M. Sandler, Sonic Visualiser: an open source application for viewing, analysing, and annotating music audio files, in: *Procs. of the ACM Multimedia 2010 International Conference*, Firenze, Italy, 2010, pp. 1467–1468.
- [42] C. Baume, Evaluation of acoustic features for music emotion recognition, in: *Procs. 134th Audio Engineering Society Convention*, Rome, Italy, 2013.
- [43] M. Ros, M. Pegalajar, M. Delgado, A. Vila, D. Anderson, J. Keller, M. Popescu, Linguistic summarization of long-term trends for understanding change in human behavior, in: *Fuzzy Systems (FUZZ)*, 2011 IEEE International Conference on, 2011, pp. 2080–2087. doi:10.1109/FUZZY.2011.6007509.
- [44] V. Novák, V. Pavliska, I. Perfilieva, M. Stepnicka, F-transform and fuzzy natural logic in time series analysis, in: J. Montero, G. Pasi, D. Ciucci (Eds.), *Procs. 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-13)*, Atlantis Press, Milano, Italy, 2013. doi:10.2991/eusflat.2013.6.
- [45] J.-J. Nattiez, *Fondements d’une sèmiologie de la musique*, Union Générale d’Editions, Paris, 1975.
- [46] A. Smaill, G. Wiggins, M. Harris, Hierarchical music representation for composition and analysis, *Computers and the Humanities* 27 (1) (1993) 7–17. doi:10.1007/BF01830712.
- [47] E. Cambouropoulos, G. Widmer, Automated motivic analysis via melodic clustering, *Journal of New Music Research* 29 (4) (2000) 303–317. doi:10.1080/09298210008565464.
- [48] O. Lartillot, Automated extraction of motivic patterns and application to the analysis of Debussy’s *Syrinx*, in: T. Klouche, T. Noll (Eds.), *Mathematics and Computation in Music*, Vol. 37 of *Communications in Computer and Information Science*, Springer Berlin Heidelberg, 2009, pp. 230–239.
- [49] G. A. Wiggins, Cue abstraction, paradigmatic analysis and information dynamics: Towards music analysis by cognitive model, *Musicae Scientiae special issue: Understanding Musical Structure and Form: papers in honour of Irène Deliège* (2010) 307–332.
- [50] J. Cohen, A coefficient of agreement for nominal cases, *Educational and Psychological Measurement* 20 (1) (1960) 37–46.
- [51] J. Landis, G. Koch, The measurement of observer agreement for categorical data, *Biometrics* 33 (1977) 159–174.