# POLYPHONIC INSTRUMENT RECOGNITION FOR EXPLORING SEMANTIC SIMILARITIES IN MUSIC

*Ferdinand Fuhrmann,*
Music Technology Group,
Universitat Pompeu Fabra
Barcelona, Spain
`ferdinand.fuhrmann@upf.edu`

*Perfecto Herrera,*
Music Technology Group,
Universitat Pompeu Fabra
Barcelona, Spain
`perfecto.herrera@upf.edu`

## ABSTRACT

Similarity is a key concept for estimating associations among a set of objects. Music similarity is usually exploited to retrieve relevant items from a dataset containing audio tracks. In this work, we approach the problem of semantic similarity between short pieces of music by analysing their instrumentations. Our aim is to label audio excerpts with the most salient instruments (e.g. *piano, human voice, drums*) and use this information to estimate a semantic relation (i.e. similarity) between them. We present 3 different methods for integrating along an audio excerpt frame-based classifier decisions to derive its instrumental content. Similarity between audio files is then determined solely by their attached labels. We evaluate our algorithm in terms of label assignment and similarity assessment, observing significant differences when comparing it to commonly used audio similarity metrics. In doing so we test on music from various genres of Western music to simulate real world scenarios.

## 1. INTRODUCTION

Music recommenders, automatic taggers, or corpus-based concatenative synthesis systems – to name just a few – use similarity measures to retrieve relevant items from an audio database (e.g. [1], [2], [3]). Here, the concept of similarity is often defined by a metric distance between low-level audio feature vectors. This distance is often used to estimate proximity of points in a highly dimensional parameter space. It has been argued in literature that both the dimensional and metric approaches are to question and that comparing many categorical and discrete features better resembles human judgments of similarity for certain stimuli [4]. In particular, similarity between pieces of music (or *music* similarity) is difficult to model with mathematical abstractions of pure acoustical relationships [5]. As a perceptual phenomenon it is defined by human auditory perception per se. In other words, no *music similarity* without *perception* [6]. Consequently, modelling of *music* similarity means addressing auditory perception and musical cognition.

Research in Music Information Research (MIR) currently abounds in examples of an observed phenomena entitled *glass ceiling*. Although state-of-the-art algorithms score around 75% of accuracy on various tasks [7], it seems nearly impossible to go beyond the current performance figures. This apparent shortcoming has been attributed to the so-called *semantic gap* which arises from loose or misleading connections between low-level descriptors of the acoustical data and high-level descriptions of the associated semantic concepts, be it in classification or in similarity assessment ([8],[9]). However, both aforementioned terms can be identified as conceptual problems, arising from the same source, namely treating a perceptual construct such as music as pure, independent in it, data corpus (i.e. ignoring its inherent qualities like social, emotional, or embodiment facets) [6]. Fig. 1 illustrates the apparent discrepancy between acoustically and semantically obtained music similarity; although the low-level information indicates a stronger correlation of track B and C, the semantic labels related to the instrumentation of all songs reveal a different similarity. Furthermore, while description or transcription of monophonic music can be roughly considered as "solved", research on many polyphonic problems is still in its infancy and the community is lacking of robust algorithms for polyphonic pitch and onset extraction, source separation, or for the extraction of higher-level concepts like chord or timbre qualities.

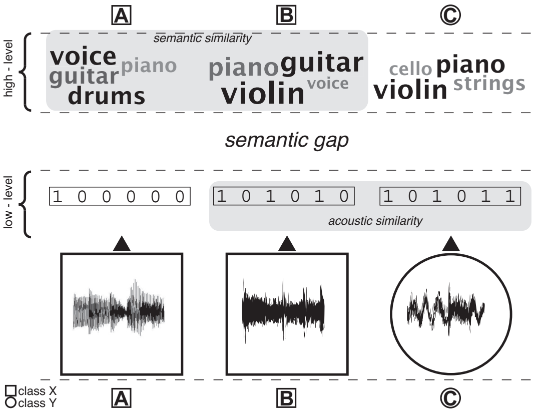

Figure 1: *The semantic gap and its roots. In this example, the low-level description of the audio content yields a different association between the tracks A, B and C than the semantic concepts related to the instruments do.*

In this work we want to automatically tag a whole audio excerpt with labels corresponding to the most relevant instruments that can be heard therein (e.g. *piano, sax, drums*), and use these labels to estimate instrument-based semantic similarities between audio files in a dataset. As the instrumentation of an audio excerpt is one of the primary cues the human mind uses to establish associations between songs (see [10] and references therein), it is directly related to music similarity and therefore human perception. Here, our focus lies on developing a general methodology for determining instrumental similarity – both in terms of the underlying data and modelled instruments. In other words, our aim is not a complete modelling of musical instruments nor any musical style one can think of – a far too ambitious goal with nowadays signal processing and MIR algorithms. Therefore our results – although not perfect – will shed light on theoretical and conceptual issues related to semantic similarity in music. Moreover, the developed similarity may be used in any music analysis, transformation, or creation system.

In the presented system polyphonic instrument classifiers are applied to tag excerpts of music. We use classifiers for 3 percussive and 11 pitched instruments (including the human voice) [11] to get a probabilistic output curve along the excerpt for each of the target instruments. We design and evaluate three strategies to process the obtained probability curves and to assign labels to the audio excerpts. Given the instrumental tags of all audio files in the dataset we then calculate pair-wise similarities between the items. Evaluation of the label assignment is finally done by calculating precision and recall metrics for multi-label classification and the presented semantic similarity is estimated as the Pearson product-moment correlation between assigned and ground truth pair-wise similarities. Thereby we both evaluate the quality of the labelling method and compare the obtained similarities to results from distance approaches usually found in MIR.

The paper is organised as follows: the next section covers related works from MIR on estimating information about the instrumentation of a piece of music. In Sec. 3 we describe the presented system along with the different labelling strategies. Sec. 4 gives insights in the used data and the experiments done to evaluate the different approaches. Finally, after a discussion, we close the article with some conclusions.

## 2. RELATED WORK

In literature, labels related to musical instruments are mainly incorporated by systems that generate social tags from audio data. In general, these algorithms use the information of the instrumentation of a piece of music along with dozens of other human-assigned semantic concepts (e.g. genre, style, mood, or even contextual information) to propagate tags throughout and/or retrieve relevant items from a music collection. Turnbull *et al.* train a probabilistic model for every semantic entry in their dataset by modelling the respective extracted audio features with a Gaussian Mixture Model (GMM) [1]. Given all models of semantic keywords the system is able to infer the probability for each keyword for an unknown piece of music or query the collection with a purely semantic input. Reported results regarding instrumental keywords yielded a precision of .27 along with a recall value of .38. Hoffman *et al.* exploit a similar path by training Codeword Bernoulli Average (CBA) models on a vector quantised representation of their music collection [12]. Again, a probability for each label can be inferred from the models for an unknown track. Besides general performance results, no detailed information about the performance on tags referring to the instrumentation of a piece of music is reported. Finally, Eck *et al.* use a music collection consisting of about 90.000 tracks from 1.277 artists to train and evaluate boosted decision stump classifiers for auto-tagging [13]. The 60 most popular tags extracted from nearly 100.00 artist in the social network Last.fm are taken for analysis, in which the categories genre, mood, and instrumentation form 77% of all labels.

Furthermore, there has been interest in the problem of identifying musical instruments from audio data. A comprehensive overview of works dealing with instrument classification from monophonies as well as polyphonies can be found in [14]. In a more recent work, Essid *et al.* developed a methodology to directly classify the instrumentation within a narrow, data-driven taxonomy [15]. Instead of modelling the musical instruments itself, classifiers were trained on the various combinations of instruments (e.g. *trumpet+sax+drums*, *sax+drums*, etc.) of the training data. The categories were derived from a hierarchical clustering, whereas the labels were manually assigned to the respective clusters. Every [16] evaluated a large corpus of audio features to discriminate between pitched sources in polyphonic music. Events containing stable pitched sources were extracted from the music pieces and features computed from the resulting excerpts. Then, clustering of the values was applied to yield a performance measure of the separability of the applied features. Recently, Heittola *et al.* presented a multi-staged system incorporating f0-estimation, source separation and instrument modelling for instrument classification from artificial mixtures [17]. A Non-negative Matrix Factorisation (NMF) algorithm is using the information provided by the pitch estimator to initialise its basis functions and to separate the sources. After separation, features are extracted from the resulting streams and classified by GMMs. Finally, Fuhrmann *et al.* trained statistical models of musical instruments with features directly extracted from polyphonic music [11]. Support Vector Machine (SVM) models for both pitched and percussive instruments were developed along with an evaluation of the temporal modelling of the used audio features.

The aforementioned works either strictly deal with instrument classification on a frame basis, i.e. the systems are built and evaluated on the correct number of instruments detected in every frame, or predict instrumental tags from a “bag-of-concepts”, where the meaningfulness of the accumulated extracted information (i.e. the musical instrument) cannot be fully assured due to limitations of the quality of user ratings and the amount of data for modelling. Please note that the here-presented approach is methodologically quite different, as it attaches a finite set of labels related only to the instrumentation to a whole audio excerpt, according to the most confident classifier decisions (e.g. “This is a piece with *flute*, *violin*, and *organ*”). To our knowledge, no study in literature approached the problem in this way.

## 3. METHOD

In this section we describe our approaches of assigning instrumental labels to audio excerpts. The front end, which is used by all three labelling methods, consists of an instrument classification system. It outputs probabilistic estimates for each of the modelled instruments on a frame basis. The so-obtained probability curves are then processed by the labelling algorithm to assign a set of labels and respective confidences to the audio excerpt.

### 3.1. Front End

Given an unknown – presumably multi-voiced – input audio excerpt, previously trained polyphonic instrument classifiers are applied within a sliding window[1]. The classifiers are trained with 11 pitched (namely *cello, clarinet, flute, acoustic and electric guitar, hammond organ, piano, saxophone, trumpet, violin*, and *human voice*) and 3 unpitched instruments from the drums set (*bassdrum*,

[1] The parameters for window length and hop size are set to 2.5 and 0.5 sec, respectively.

*snaredrum*, and *hihat*). The training data for the pitched instruments consist of $2.5$ seconds long polyphonic segments containing predominant target instruments, all data taken from commercially available music[2]. Percussive instruments are trained with $.15$ sec excerpts extracted from data of two public datasets, namely the ENST [18] and MAMI [19] collections. Typical audio features representing timbre were extracted frame-wise and integrated over the segment length using mean and variance statistics of the instantaneous and delta values to train the instrumental models (see [11] for more details). The classifiers – we used support vector machines (SVMs) – output probabilistic estimates for all the mentioned instruments which leads to 14 probability curves along the segment.

### 3.2. Labelling

In the following we describe the methodology we have taken to integrate the classifiers' decisions to yield the final set of labels and respective confidences for a given audio excerpt.

Contrary to the processing of the pitched instruments, where we are interested in assigning a possible label for all the modelled instruments, we simplify the labelling of the percussive instruments. Here, we accumulate the three probability curves (i.e. *bassdrum*, *snaredrum* and *hihat*) to label the excerpt with either *drums* or *no-drums*. Similar to the *Percussion Index* presented in [20], we count the number of unlabelled onsets and divide it by the total number of onsets[3], given the estimated onsets inside the audio[4]. If this ratio exceeds an experimentally defined threshold $\theta_{ratio}$, the excerpt is labelled with *no-drums*, otherwise with *drums*.

For the labelling of pitched instruments, we process all probability curves which hold a mean probability value along the segment greater than the activation threshold $\theta_{act}$. Furthermore, to filter out unreliable excerpts, we define an uncertainty area determined by the upper and lower values $\theta_{up}$ and $\theta_{lo}$: if the 3 highest mean probability curves fall into this area (as it signals the absence of discriminable instruments) the excerpt is skipped and not labelled at all. This is motivated by experimental evidence as, on excerpts with heavy inter-instrument occlusion or a high number of not modelled instruments, the classifier output shows this typical behaviour. With the remaining probability curves we then examine three different strategies for labelling:

**Mean Probability Values (MPV)** Labelling is simply done by taking the highest $n_{MPV}$ mean probability instruments. The respective label confidences are set to the mean probability values of the instruments. Following this approach, temporal information is completely disregarded, as all probabilities are averaged along the excerpt.

**Random Segment Selection (RSS)** Random segments of length $l_{RSS}$ are taken from the audio input to account for variation in the instrumentation. Within each of these segments, a majority vote is performed to attach either one or – in the case of a draw – two labels to the random segment. The assigned confidences are a result of the number of the majority label(s) divided by both the length $l_{RSS}$ and the total number of random segments extracted from the input. All labels from the $n_{RSS}$ random segments are merged and the confidences of multiple instances assigned to the same label summed.

**Curve Tracking (CT)** Probably the most elaborate and plausible approach from the perception point-of-view: classification is done in regions of the audio excerpt where a dominant instrument can be clearly identified. Decisions in regions where overlapping components hinder confident estimations are inferred from context. Therefore, we scan all instrument probability curves for piece-wise predominant instruments. Here we define predominance as having the highest probability value for 90% of a segment with minimum length $l_{CT}$. Once a predominant instrument is located, its label is attached to the audio excerpt along with a confidence defined by the ratio of the found segment's length to the total length of the excerpt. This process is repeated until all regions with predominant instruments are found. Finally, all labels are merged and multiple confidences of the same label added. During this process, we explicitly use the temporal dimension of the music itself (and thereby the contextual information provided by the classifiers' decisions) to infer a set of labels.

Given the set of labels and their respective probabilities for an audio excerpt, a final threshold $\theta_{lab}$ is used to filter out labels which hold a too low probability value.

## 4. EXPERIMENTS

### 4.1. Data

For our experiments we collected a total number of 100 pieces of Western music, spanning a diversity of musical genres and instrumentations. It should be noted that the musical data for training the polyphonic instrument classifiers and the data for the current experiments were taken from different sources[5]. Two subjects were paid for annotating a half of the collection each. After completion, the data was swapped among the subjects in order to double-check the annotation. Moreover, a third person reviewed all the so-generated annotations. In particular, the on- and offsets of nearly all instruments were marked manually in every file, whereas no constraints in the vocabulary size were imposed. This means that, in addition to the labels of the 11 modelled instruments and the label *drums*, every instrument was marked with its corresponding name. Hence, the number of categories in the test corpus is greater than the number of categories modelled by the instrument classifiers. Moreover, if an instrument was not recognised by the subject doing the manual annotation, the label *unknown* was used.

For all following experiments we split the data into a development and a testing set by assigning $1/3$ of the corpus to the former and the rest to the latter subset. Table 1 shows the genre distribution of the whole 100 tracks and Fig. 2 and 3 show the frequency of all annotated instruments and the number of instruments annotated per track, respectively. We hypothesise that with increasing number of tracks the shape of the histogram in Fig. 3 will resemble a gaussian distribution with its mean between 4 and 6 annotated

[2] In total, this training collection covers more than 2.500 pieces of music to account for the noise introduced by the underlying polyphony.

[3] we count an onset as unlabelled if none of the three probability values at the respective onset exceeds the threshold $\theta_{dru}$.

[4] we used an energy based onset detection algorithm [21] to infer the drum onsets.

[5] This means that it is impossible that a certain piece of music appears in both datasets. Moreover, within each collection there are no two tracks of the same artist to avoid the so-called *Artist* and *Album effects*.

Table 1: *Number of tracks with respect to the different musical genres covered by the whole dataset.*

| **rock** | **pop** | **classic** | **jazz** | **electronic** | **folk** |
|---|---|---|---|---|---|
| 25 | 23 | 13 | 22 | 8 | 9 |

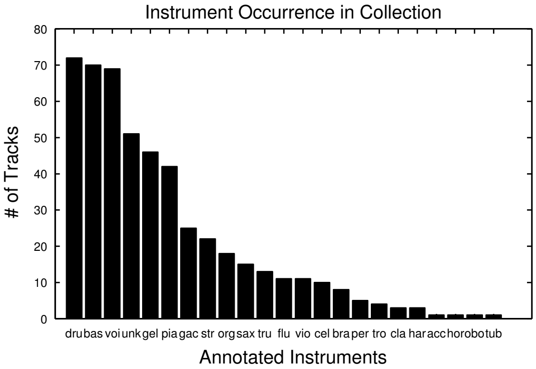

Figure 2: *Frequency of annotated instruments in the used music collection. Please note that all instruments modelled by the polyphonic recognition modules are top ranked.*

instruments. Additionally, for estimating the proportion of instruments not modelled by the classifiers, we compute the mean ratio of modelled-to-total labels in a track (.71) along with the average number of not-modelled instruments per track (1.61).

### 4.2. Labels

Besides the 11 modelled pitched instruments and the already mentioned "fused" label *drums*, we introduce the two composite labels *bra* (for brass sections) and *str* (for string ensembles) for evaluation purposes. This is motivated by the fact that both are frequent labels used to describe the instrumentations of a given piece of music (see Fig. 2). As they are not modelled by the polyphonic instrument classifiers, the individual predictions have to be adapted depending on the respective set of ground truth labels. We therefore substitute every *cel* and *vio* in the set of predicted labels with *str* whenever there is a label "strings" in the annotation. Similarly, we process the labels *cla*, *sax*, and *tru* when we find a *bra* in the respective set of ground truth labels.

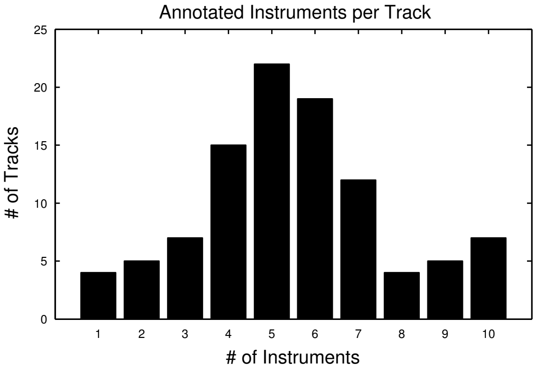

Figure 3: *Histogram of the number of instruments annotated per music track in the used collection.*

### 4.3. Metrics

In this section, we introduce the metrics used to evaluate the different algorithms presented in the paper. First we define several metrics to estimate the performance of the instrumental tag assignment given the ground truth annotations. Then we present a measure of semantic similarity between two items, which have been labelled by the aforementioned tagging algorithm.

#### *4.3.1. Labelling*

For estimating the labelling performance, the underlying problem to evaluate is multi-class multi-label classification. Please note that in our specific case, as there has not been any restriction in the vocabulary size for the manual annotations, the set of all labels $L$ in the dataset is theoretically not closed. But when considering only those labels, which are actually used to describe the instrumental content of an audio excerpt (i.e. the 11 modelled pitched instruments, *drums*, and the two composite labels *brass* and *strings*), we can regard it as closed without loss of generality.

Consider $L$ the closed set of labels $L = \{l_i\}, i = 1 \dots N$. Given the audio dataset $X = \{x_i\}, i = 1 \dots M$, with $M$ items, $\hat{Y} = \{\hat{y}_i\}, i = 1 \dots M$, the set of ground truth labels for each $x$ and $Y = \{y_i\}, i = 1 \dots M$, and $y_i \subseteq L$, the set of predicted labels assigned to the audio excerpts in $X$. We then define precision, recall, and F-measure for every label in $L$:

$$P_l = \frac{\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{i=1}^{M} y_{l,i}}, \quad \text{and} \quad R_l = \frac{\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{i=1}^{M} \hat{y}_{l,i}}, \tag{1}$$

$$F_l = \frac{2P_l R_l}{P_l + R_l}, \tag{2}$$

where, for any given instance $i$, $y_{l,i}$ and $\hat{y}_{l,i}$ denote boolean variables indicating the presence of label $l$ in the set of predicted labels and in the ground truth annotation, respectively. Furthermore, to introduce a general performance metric, we define the unweighted mean of label F-measures as

$$F_{macro} = \frac{1}{|L|} \sum_{l=1}^{|L|} \frac{2\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{i=1}^{M} y_{l,i} + \sum_{i=1}^{M} \hat{y}_{l,i}}, \tag{3}$$

where $|L|$ denotes the cardinality of $L$. As $F_{macro}$ does not account for individual label distributions (i.e. less frequent labels contribute the same amount to the metric as more frequent ones do), we additionally introduce

$$F_{micro} = \frac{2\sum_{l=1}^{|L|}\sum_{i=1}^{M} y_{l,i}\hat{y}_{l,i}}{\sum_{l=1}^{|L|}\sum_{i=1}^{M} y_{l,i} + \sum_{l=1}^{|L|}\sum_{i=1}^{M} \hat{y}_{l,i}}, \tag{4}$$

which considers the predictions for all instances together.

Although the presented F-measure metrics give an objective and adequate performance measure, under certain circumstances it is of advantage to evaluate the general system performance with precision and recall measures separately. We therefore define

$$Pre = \frac{1}{\sum_{l=1}^{|L|}\sum_{i=1}^{M} y_{l,i}} \sum_{l=1}^{|L|} (\sum_{i=1}^{M} y_{l,i}) P_l, \qquad (5)$$

and

$$Rec = \frac{1}{\sum_{l=1}^{|L|}\sum_{i=1}^{M} \hat{y}_{l,i}} \sum_{l=1}^{|L|} (\sum_{i=1}^{M} \hat{y}_{l,i}) R_l, \qquad (6)$$

the weighted mean precision and recall across all labels, respectively.

### *4.3.2. Similarity*

We then introduce a measure of music similarity using the semantic descriptions attached to the audio tracks (i.e. the instrumental tags). Instead of using a geometric model, which has been proven to be problematic under certain assumptions (see e.g. [4] and references therein for details), we apply metrics from set-theory to estimate associations based on the instrumentation between the audio files in our dataset.

Again, assume $X = \{x_i\}, i = 1 \ldots M$ being a set objects, each represented by a set of labels $y \in Y$. We then define $s(x_i, x_j)$ to be a measure of similarity between $x_i$ and $x_j$, for all $x_i, x_j \in X$, given the matching function $F$ [4]:

$$s(x_i, x_j) = F(y_i \cap y_j, y_i - y_j, y_j - y_i), \qquad (7)$$

that is, the similarity between $x_i$ and $x_j$ is expressed by a function of their common and distinct labels. Following [4], we finally define a similarity scale $S$ and a non-negative scale $f$ such that for all $x_i, x_j \in X$,

$$S(x_i, x_j) = \frac{f(y_i \cap y_j)}{f(y_i \cap y_j) + \alpha f(y_i - y_j) + \beta f(y_j - y_i)}, \qquad (8)$$

for $\alpha, \beta \geq 0$. This relation, also known as *ratio model*, normalises similarity so that $S$ is between 0 and 1.

## 4.4. Parameter tuning

The development set is used to find the optimal parameter values yielding the best overall labelling performance of the algorithm. We evaluate a grid search over a predefined discrete value range for each relevant parameter. The best values are then determined by the top scoring $F_{micro}$ values[6]. Table 2 shows parameter acronyms, predefined discrete values set, and best found values, respectively.

Table 2: *Acronyms and respective discrete values of the parameters used in the grid search for training. Bold values indicate best performance among tested values. See Sec. 3.2 for the exact parameter meanings.*

| Acronym | Value |
|---|---|
| $\theta_{act}$ | [.09, .14, **.18**, .27, .45, .68] |
| $\theta_{up}$ | [.14, .18, **.27**] |
| $\theta_{lo}$ | .09 |
| $\theta_{lab}$ | [.05, .1, **.2**, .3] |
| $\theta_{dru}$ | [**.5**, .6, .7, .8] |
| $\theta_{ratio}$ | [.3, **.5**, .7] |
| $n_{MPV}$ | [1, **2**, 3, 4] |
| $l_{RSS}$ | [**2**, 3, 4, 5] *(sec.)* |
| $n_{RSS}$ | $max.\ 4$ |
| $l_{CT}$ | [**2.5**, 3.5, 4.5, 5.5] *(sec.)* |

## 4.5. Labelling evaluation

### *4.5.1. Preprocessing*

To obtain excerpts for experimental analysis, we segment the pieces of music in the test set using a Bayesian Information Criterion (BIC) segmentation algorithm [22]. This unsupervised algorithm, working on frame-wise extracted features, is used to find changes in the time series of the input data. We use the first 13 Mel Frequency Cepstral Coefficients (MFCCs) [23], extracted from 40 Mel bands, as input to the algorithm, accounting for the timbre and its changes along the track. The algorithm shifts a texture window along the audio, which is split into two parts, where the whole content of the window and its subparts are fit to a specific model[7]. The BIC-value – in general defined by the maximum log-likelihood ratio of a given model and a penalty term – is then calculated by the difference of the maximum likelihood ratio test (determined by the covariance matrices of the three models) and the penalty term. If this value exceeds a certain threshold, a change point is detected, and the window is shifted for the next analysis. We refer to [24] for details on the implementation. Besides, with the corresponding parameter settings the algorithm can also be used to find boundaries between structural blocks of a song.

We segment all songs of our test collection using the aforementioned algorithm. If possible, we then take the 4 longest segments of each track to build the final test set, yielding a total amount of 255 audio excerpts.

### *4.5.2. Results*

In order to compare our results to a chance baseline, we introduce a random label assignment algorithm. It assigns a number of labels with corresponding confidences to each of the generated excerpts. The number of labels and the corresponding confidences are taken randomly from the distribution of the number of labels and of confidence values, respectively[8]. The former is modelled as a histogram whereas the latter correspond to a normal distribution $\mathcal{N}(\mu, \sigma)$ with mean $\mu$ and standard deviation $\sigma$, whereas both distributions are determined by the observed data. The label itself is randomly drawn from the distribution of annotated labels in the test set.

We now present the results obtained for each of the labelling methods, including the respective means of 10 runs of the random label assignment, by evaluating the attached tags against the ground truth annotations. An analysis of variance of instance $F_{micro}$ values shows no significance for pair-wise comparison of the three

[6] It should be noted that it is only our convention that the best parameter values correspond to the highest $F_{micro}$ score. Depending on the application and its needs, another metric (e.g. precision) could define the best overall labelling performance and serve a different set of best parameter values.

[7] here, the data is fit to a single gaussian distribution.

[8] The distributions are obtained when processing the test collection with the $CT$ labelling method and its best parameter settings from Table 2.

Table 3: *Evaluation results for tag assignment on the testing data. We used the respective optimal parameters depicted in Table 2 for each of the 3 labelling methods. The random method values correspond to the mean of 10 independent runs.*

| Method | $Pre$ | $Rec$ | $F_{macro}$ | $F_{micro}$ |
|---|---|---|---|---|
| **rand** | 0.424 | 0.155 | 0.11 | 0.227 |
| **MPV** | 0.86 | 0.534 | 0.441 | 0.659 |
| **RSS** | 0.857 | 0.521 | 0.429 | 0.648 |
| **CT** | 0.864 | 0.528 | 0.442 | 0.656 |

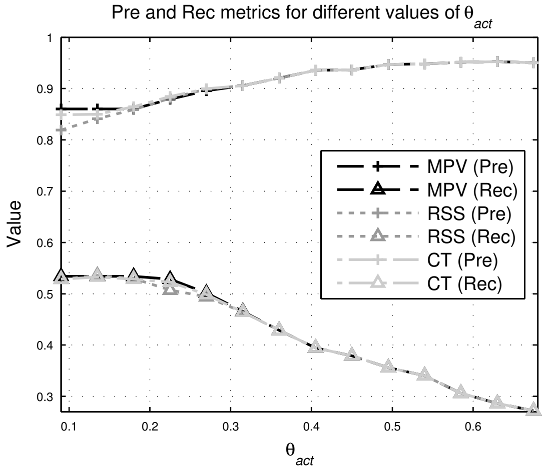

Figure 4: *Precision and recall metrics for varying values of $\theta_{act}$. As it can be seen $\theta_{act}$ determines the sensitivity of the labelling algorithm: depending on its value the labelling performance metrics show very different outputs.*

labelling methods $MPV$, $RSS$, and $CT$. However, the average instance $F_{micro}$ value of the combined three methods ($M = 0.31$, $SD = 0.15$) was significantly higher than the one of the random label assignment ($M = 0.12$, $SD = 0.13$), $F(1, 508) = 217.29$, $p < .001$. Table 3 shows the evaluation metric values for the respective best parameter settings found in the training.

Additionally, Fig. 4 shows the precision and recall metrics $Pre$ and $Rec$ for different values of $\theta_{act}$. For each labelling method we used the respective best parameter settings from Table 2. Finally, the system performance in correctly identifying individual instrument categories for all labelling methods is depicted in Fig. 5.

### 4.6. Similarity Assessment

Using the instrumental tags assigned to the audio excerpts in our dataset we then compute pair-wise similarities between the tracks. In accordance with Eq. (8), we need to determine three parameters: the scale $f$, measure of the common and distinct features, and the parameters $\alpha$ and $\beta$, which weight the influence of the respective distinct features to each other.

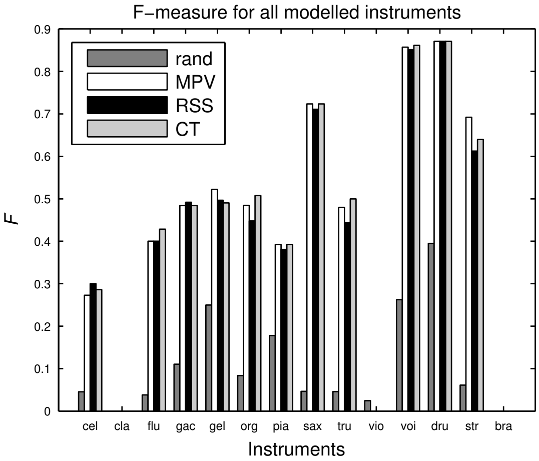

Figure 5: *F-measures for individual instruments. $F$ values are plotted for all labelling algorithms, including the random assignment.*

The parameters $\alpha$ and $\beta$ define the symmetric aspects of the similarity measure. Suppose any non-symmetric similarity relation $S(a, b)$, where the labels of $a$ have more weight than the labels of $b$. By setting $\alpha > \beta$, the distinct labels of $a$ get a higher weight as the ones of $b$, thus contributing more to the overall similarity measure. However, the problem here can be regarded as symmetric (i.e. $S(x_i, x_j) = S(x_j, x_i)$). We therefore set the parameters to $\alpha = \beta = 1/2$, reducing Eq. (8) to

$$S(x_i, x_j) = \frac{2f(y_i \cap y_j)}{f(y_i) + f(y_j)}. \tag{9}$$

Finally, the scale $f$ has to be determined. One straightforward approach would be to simply use the counting measure. Thus, similarity is estimated by just counting the number of common and distinct features. As it obviously puts the same weight to every label regardless of its frequency in our dataset, we weight each label by its relative occurrence in the dataset before summing.

A proper evaluation of the obtained pair-wise distances would require ground truth data based on similarity ratings from human listeners. Although desirable, these are not available in the current stage of the research process and therefore remain out of the scope of this work. However, we can relate our observed data to results from previously used distance approaches. Therefore, we first build binary feature vectors from the assigned instrumental labels and calculate the pair-wise euclidean distances between them. Second, we model each audio excerpt in our test set as a single gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$ (both diagonal and full covariance matrices are considered) based on frame-wise extracted MFCCs. The distance between two models is then expressed by the symmetric Kullback-Leibler divergence. This approach has been shown to be superior in similarity problems where timbral information is pivotal (i.e. artist and album similarities)[25].

In order to estimate how well the results resemble the semantic similarity expressed by Eq. (9), we correlate the observed pairwise distances – obtained by both the semantic and euclidean distance approach using the computed instrumental labels, as well as the gaussian modelling via the Kullback-Leibler divergence – with the similarities obtained by applying Eq. (9) to the manual annotated labels. Table 4 shows the resulting Pearson product-moment correlation coefficients.

Table 4: *Pearson product-moment correlation coefficients for the four similarity test scenarios. The first column represents the similarity obtained via Eq. (9), the second the euclidean distances from the instrumental tags, the third and forth the distances resulting from the gaussian modelling with diagonal and full covariance matrix, respectively. All obtained correlations hold significance values* $p < .001$.

| $semantic$ | $euclidean$ | $KL_{diag}$ | $KL_{full}$ |
|---|---|---|---|
| 0.54 | −0.24 | −0.11 | −0.12 |

## 5. DISCUSSION

The results presented in the precedent sections demonstrate the capabilities and potentials of our algorithm and therefore substantiate our taken methodologies. On the one hand it is shown that with a standard pattern recognition approach towards musical instrument modelling in polyphonies, and with a straightforward and simple labelling strategy, reliable tags containing information about the instruments playing can be attached to an audio excerpt, regardless its musical genre or instrumental complexity. Moreover, these labels can be used to construct basic and effective associations between audio tracks, based on their semantic relations concerning the instrumentation. On the other hand, much room for improvements can be identified, both in classification and labelling. We will now discuss all parts of our algorithm consecutively:

First let us examine the polyphonic instrument classification. Given the fact that there are still 8 categories in the ground truth annotations which are not modelled by the classifiers, we see some need in adapting the instrumental modelling in this regard (see Fig. 2). Moreover, the category *unknown* is ranked on 4th position, indicating that we are still lacking the right concept to overcome problems with inputs which are not known by the system[9]. A simple solution regarding the unknown categories would be to move away from predicting the presence of the instrument playing towards a more general concept of *this instrument sounds like...*. However, the predictions for the trained instruments are robust and are shown to be useful in our context.

Regarding the labelling methods we can observe that none of the proposed methods performs superior than the others. This is even more surprising when considering the conceptual difference of taking just the mean probability of the instruments along the whole segment and scanning their output probabilities for piecewise maxima. We may explain it by the fact that if an instrument is predominant it is recognised by all three methods without problems. On the other hand, if the algorithm is faced with an ambiguous scenario, all methods perform equally bad.

When looking at the instrument-specific performance of the labelling algorithm, we can observe an excellent performance with the labels *drums* and *voice*. Also the labelling of the instruments *sax*, *organ*, *trumpet*, *acoustic guitar*, and *electric guitar* as well as the composite label *strings* yield satisfactory results of our evaluation metrics. The *piano* performs slightly inferior as the aforementioned, but it is not clear if the resulting $F$ value in Fig. 5 is due to a low precision or recall. We hypothesise that as the piano is often used as an accompaniment instrument for the human voice, the value is due to a low recall. Moreover, the low performance of the *violin* can be explained by the merging of the labels when creating the composite label *strings* (i.e. the label *vio* mostly appears together with the label *str*, and therefore all predictions of *vio* are transformed into predictions of *str*). Furthermore, *cla* and *bra* only appear in a minority of the audio excerpts under analysis.

Reviewing the different parameters in Table 2 and their impacts on the overall labelling performance, $\theta_{act}$ is the most influential one. Of course, small adjustments in performance can also be accomplished by varying $\theta_{lab}$, $l_{RSS}$, or $l_{CT}$ , but $\theta_{act}$ determines the overall sensitivity of the algorithm. Depending on the need of the application using the instrumental tagging algorithm, one can adjust the number of true and false positives by simply altering this parameter (see Fig. 4).

Nonetheless, in general the labelling algorithm is only able to identify a fraction of all instruments playing in an audio excerpt. This is due to the fact that primarily predominant sources are identified. On average, the algorithm outputs 2 labels per excerpt, which is less than half of the maximum that can be observed in Fig. 3[10]. Evidently we will not be able to recognise instruments in a dense mixture without more elaborate signal processing tools like source enhancement or polyphonic pitch and onset detection. Moreover, to improve recognition performance we clearly identify a need for a complete probabilistic modelling with knowledge integration from different sources. Also, prior information could be very useful (e.g. reliable genre information can reduce the number of instruments to recognise, thus minimising the error introduced by instrument confusions). However, deploying the information of the predominant instruments is not only useful for transformation and computational analysis, but also important from the perceptual point-of-view, as the predominant sources contribute most to the overall timbral sensation of the audio excerpt.

Regarding the presented semantic similarity, the used measure is both simple and intuitive. Our approach, which is solely based on the overlap of the predicted labels, resembles ground truth similarities and shows significant differences when compared to a distance approach applied to the tags as well as to metric-based approaches based on low-level features. From the results presented in Table 4 there is evidence to suggest that it reflects both cognitive principles and carries complementary information with respect to the other similarity estimations. On the other side, the similarity we are presenting relies on a simple merging of instrumental labels along the segment to form a closed set. It remains more than to question if this merging resembles similarity judgments of humans based on timbre. Moreover, in what extent instrumental information is used by humans to find associations between pieces of music is difficult to estimate, but this information may serve as an essential brick in the concept of a general audio similarity.

In general, the presented method is thought to be used in music creation, transformation and analysis algorithms. When retrieving relevant items from a database, the concept of relevance can be extended by the presented instrumental similarity. It may add an interesting aspect to these systems which largely rely on similarity metrics based on geometric models. Or consider any music modelling algorithm, be it for genre classification, for mood estimation or, more general, for similarity assessment; having an idea about the instrumentation of the analysed track can dramatically reduce the parameter space to search for and, therefore, lead to more ro-

[9] Besides, this problem is prototypical for many classification tasks and only a minority of works are considering it as part of their approach.

[10] Please recall that we are only tagging excerpts taken from full pieces of music. The problem may be reduced when analysing different segments of one track and combining the so found labels.

bust – thus perceptually more plausible – results.

## 6. CONCLUSIONS

In this article a general methodology to derive a semantic similarity based on the instrumentation of an audio excerpt was presented. We used polyphonic instrument classifiers to process segments of music and integrate their predictions over the whole excerpt. On this basis, three strategies for assigning tags corresponding to the instrumentation were examined. Thereby we did not find any superior method, indicating that labelling performance is not dependent on the specific method. Furthermore, we introduced a measure of similarity coming from set-theory, which is only based on label overlap, and is rooted on the way humans judge conceptual similarities. Labelling performance evaluation yielded precision values up to 0.86 and F-measures greater than 0.65 (for random baselines of 0.41 and 0.22, respectively); moreover, significant differences were observed when comparing the presented similarity estimation with metrics usually found in MIR systems. The developed algorithm may be used in any music creation, transformation, or analysis system.

## 7. ACKNOWLEDGEMENTS

This research has been partially funded by the Spanish Ministry of Industry, Tourism and Trade project CLASSICAL PLANET.COM TSI-070100-2009-407[11].

## 8. REFERENCES

[1] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio Speech and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[2] M. Sordo, C. Laurier, and O. Celma, "Annotating music collections: How content-based similarity helps to propagate labels," *Proc. of ISMIR*, pp. 531–534, 2007.

[3] D. Schwarz, "A system for data-driven concatenative sound synthesis," *Proc. of DAFx*, pp. 97–102, 2000.

[4] A. Tversky, "Features of similarity," *Psychological review*, vol. 84, no. 4, pp. 327–352, 1977.

[5] J. Aucouturier, "Sounds like teen spirit: Computational insights into the grounding of everyday musical terms," *Language, Evolution and the Brain*, pp. 35–64, 2009.

[6] G. Wiggins, "Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music," *Proc. of IEEE ISM*, pp. 477–482, Oct 2009.

[7] S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[8] J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?," *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, Jan 2004.

[9] O. Celma and X. Serra, "Foafing the music: Bridging the semantic gap in music recommendation," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 4, pp. 250–256, 2008.

[10] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Perception*, vol. 27, no. 3, pp. 223–241, 2009.

[11] F. Fuhrmann, M. Haro, and P. Herrera, "Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music," *Proc. of ISMIR*, 2009.

[12] M. Hoffman, D. Blei, and P. Cook, "Easy as CBA: A simple probabilistic model for tagging music," *Proc. of ISMIR*, pp. 369–374, 2009.

[13] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic generation of social tags for music recommendation," *Advances in neural Information Processing Systems*, vol. 20, 2007.

[14] P. Herrera, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," *Signal Processing Methods for Music Transcription*, pp. 163–200, 2006.

[15] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 68–80, 2006.

[16] M. Every, "Discriminating between pitched sources in music audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 267–277, 2008.

[17] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," *Proc. of ISMIR*, 2009.

[18] O. Gillet and G. Richard, "ENST-drums: an extensive audio-visual database for drum," in *Proc. of ISMIR*, 2006.

[19] K. Tanghe, M. Lesaffre, S. Degroeve, M. Leman, B. De Baets, and J. Martens, "Collecting ground truth annotations for drum detection in polyphonic music," in *Proc. of ISMIR*, 2005.

[20] M. Haro and P. Herrera, "From low-level to song-level percussion descriptors of polyphonic music," *Proc. of ISMIR*, 2009.

[21] P. Brossier, "Automatic annotation of musical audio for interactive applications," *Doctoral Dissertation, Centre for Digital Music, Queen Mary University of London*, 2007.

[22] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proc. of the DARPA Broadcast News Transcription & Understanding Workshop*, pp. 127–132, Nov 1998.

[23] B. Logan, "Mel frequency cepstral coefficients for music modeling," *Proc. of ISMIR*, 2000.

[24] X. Janer, "A BIC-based approach to singer identification," *Master of Science Thesis, Universitat Pompeu Fabra*, 2007.

[25] D. Bogdanov, J. Serra, N. Wack, and P. Herrera, "From low-level to high-level: Comparative study of music similarity measures," *Proc. of IEEE ISM*, pp. 453–458, 2009.

[11] `http://www.classicalplanet.com`