Novel Method of Nonlinear Symbolic Dynamics for Semantic Analysis of Auditory Scenes

Pauline Mouawad  
Laboratoire Bordelais de Recherche en Informatique  
University of Bordeaux  
Bordeaux, France  
pauline.mouawad@u-bordeaux.fr

Shlomo Dubnov  
Center for Research in Entertainment and Learning  
University of California San Diego  
California, USA  
sdubnov@ucsd.edu

Abstract— Discovering semantic information from complex signals is a task concerned with connecting humans' perceptions and/or intentions with the signal’s content. In the case of audio textures from environmental sounds produced by a cheering crowd, laughter, crackling fire, car crash or explosion, complex meanings of the events are inferred and appraised in a listener's mind, which triggers an affective response that is relevant for well-being and survival. In this paper we contribute to the ongoing research on affective semantics from sound by proposing a novel learning framework of affective auditory scene analysis using a recently developed method of non-linear dynamic signal analysis. Using an adaptive symbolization process that finds the best audio structure representation in terms of the symbolized sequence recurrence properties, we show that measures of periodicity and complexity derived from our model are relevant for the characterization of affect in auditory scenes, and that they will perform better than state-of-the-art methods relying on low-level acoustic features.

Keywords — Semantic analysis, nonlinear dynamics systems, recurrence quantification analysis, recurrence plots, symbolization, Variable Markov Oracle.

I. INTRODUCTION

Research on affect recognition in audio signals has mainly focused on mapping acoustics to higher-level affective semantics. However, capturing important emotional meaning from audio structure while disregarding trivial or irrelevant information is a complex process that cannot be determined by the stationary statistics of the low-level acoustics. In order to derive such semantic information, it is necessary to identify the structure of audio content that consists of repeating patterns that evolve according to a temporal order.

Studying the dynamical aspects of emotion in sound is gaining increased attention as researchers apply nonlinear dynamics approaches to study the role of the signal's dynamical structure and its derived complexity features in influencing various emotional perceptions. Nonlinear dynamical approaches provide a quantitative description of such fine-grained patterns that characterize semantic events.

Recent advances in research on sound dynamics include modeling various voice pathologies and vocal cord disorders from speech signals [16—18], [43]; environmental sound recognition [28, 29]; and discrimination between different singing styles [21]. Despite these advances, very few researches have applied nonlinear dynamics for modeling emotion in audio signals. In [30] measures of the geometrical properties of the phase space reconstruction (PSR) are employed for emotion recognition in speech; in [31] recurrence properties of the vowel are found to accurately describe the dynamic behavior of six basic emotions.

In this paper we contribute to the ongoing research on affective semantics from sound by proposing a novel learning framework of affective auditory scene analysis, developed in two steps. First, we use a recently developed method of nonlinear dynamic signal analysis through an adaptive symbolization process that finds the best audio structure representation in terms of the symbolized sequence recurrence properties. Our symbolization method, the Variable Markov Oracle (VMO), is a variable length memory model that can be applied to general time series, is capable of finding repeating patterns while maintaining their temporal order, and was used for gesture recognition and motif recognition in music, as well as generation of variations for machine improvisation. Moreover, one of the interesting properties of the VMO method is its ability to detect perceptual recurrences of subsequences of variable length. Second, we construct a symbolic recurrence plot (RP) from the repetitions found by the VMO, and apply our RQA analysis to sequences of symbolized audio features to uncover the meaning of the underlying dynamics with respect to affective descriptions. The contribution of our method over previous recurrence analysis methods is that our model captures the best recurrent patterns in the signal, and then RQA features explore the dynamics of these recurrences. We hypothesize that the measures of periodicity and complexity derived from our model are relevant for the characterization of affect in auditory scenes, and that they will perform better than state-of-the-art methods relying on low-level acoustic features.
I. THEORETICAL BACKGROUND

A. Nonlinear Time Series Analysis

In order to learn about the underlying dynamics of time-ordered data such as auditory sounds, it is necessary to reconstruct the state space, which is the set of possible states a system can take. Given a signal, nonlinear time series analysis reconstructs the hidden dynamics through the study of phase space trajectories that represent its dynamics over time. Considering that the system's states are often unknown, Takens time-delay embedding theorem is used to reconstruct the full dynamics of one generic signal from a single time series [3, 25].

B. Recurrence Plots

A recurrence plot (RP) is a graphical representation of a squared matrix with black and white dots at two time axes, highlighting recurrent states in the structural dynamics underlying the signal. Each black dot corresponds to a value of 1 in the matrix entry Rij and indicate that two states at different times are similar, i.e., the points x(t) and x(t) in phase space fall inside a ball of radius ε, the threshold distance [25].

C. Recurrence Quantification Analysis

RQA is a nonlinear statistical technique that explores the dynamical structure of a time series by quantifying the structures obtained in recurrence plots [3], [26], [38]. It can be applied to non-stationary processes in continuous or discrete time series. For example, the RQA metric Determinism can discriminate signals from noise, and is valuable in pattern mining and classification tasks.

Seven RQA measures are computed in this work and they are divided in two classes:

Measures based on recurrence density:

Recurrence rate (REC): is the percentage of points in the RP:

\[
REC = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}(l,j) \tag{1}
\]

Measures based on the distribution \(P(l)\) of lengths \(l\) of diagonals:

Determinism (DET): is the percentage of points in the diagonal lines:

\[
DET = \frac{\sum_{i=l_{min}}^{N} I \times P(l)}{\sum_{i,j}^{N} R_{i,j}(l,j)} \tag{2}
\]

where:

\[I = \begin{cases} 1 & \text{if } P(l) \neq 0 \\ 0 & \text{otherwise} \end{cases} \]

\[L = \frac{\sum_{l=l_{min}}^{N} P(l)}{\sum_{l}^{N} P(l)} \tag{3}
\]

D. Feature Extraction

So far the discussion of recurrence analysis assumed, even if implicitly, that the phase space of the dynamical system is directly related to the measurements themselves, which in the case of audio signal comprises of samples of the sound waveform. Such an assumption is common in non-linear dynamical system modeling in view of Takens theorem. A natural question that thus arises is whether the measured variables can indeed reconstruct the state space. In our case the situation is complicated by several factors. First, it is not clear what the underlying system is and whether the audio signal data indeed contains sufficient information about the system dynamics. It is also possible that even if all the required information is present in the audio samples, that such measurements are very noisy representations of the relevant system variables. It is known that the choice of the variable may deeply affect the quality of the reconstructed space and there might be some certain representation that provides a better observability of the original state space than others. Our attempts to use embedding directly on the signal data were not successful, mainly due to computational difficulties, such as highly increased computation time and considerable memory consumption for long time series of audio data.

Standard ways to consider similarity in audio signals is through time-frequency representation. Accordingly, in our experiments we apply dynamical non-linear analysis to a generic time-frequency representation in term of a constant-Q filter bank (CQT). Such logarithmic spacing of filter center frequencies versus bandwidths is more natural in terms of approximating human auditory analysis.

II. OUR APPROACH

One key concern when using RPs is finding the threshold to make sure that the RP exhibits enough recurrence points. Another difficulty to address is the length of the sequence used to generate the RP. This is considered as a second embedding step that is different from the phase space embedding, however in traditional RP construction methods these two steps are indistinguishable, as the RP is constructed first and then the repeated motifs are found by looking for diagonal lines.
In this work we propose a novel method that does not involve a phase space reconstruction with embedding, but instead, we perform a symbolization step by clustering the signal's feature frames with relatively large size and we transform them to logarithmically spaced frequency space using CQT. This is done using the Variable Markov Oracle (VMO) [34], a suffix automaton that reduces a multivariate time series down to a symbolic sequence while retaining the repeating subsequences. Accordingly, we consider recurrence of symbolic sequences without a need to estimate a threshold, since this step is implicitly done during the symbolization, based on a mutual information criterion that estimates the optimal threshold in terms of maximizing Information Rate (IR) [6] that considers mutual information between past and present in a signal.

The advantage of our method is that it detects the best repetition structure by searching over all possible symbolized sequences and selects the threshold that achieves the best compression effect, that is that gives the most informative symbolization first. Then it represents the structure of repeated motifs of variable length as RP. In the next section we describe this approach.

A. The Variable Markov Model

VMO [39—42] is a data structure derived from Factor Oracle (FO), a compressed suffix tree that allows a fast indexing and retrieval of repeated substrings (factors) and patterns (repeated suffixes) from a symbolic sequence, and Audio Oracle (AO) [7], a signal extension of FO, that extends the domain of FO from symbolic sequences to multivariate time series such as an audio signal, sampled at discrete times. VMO stores the information regarding repeating sub-sequences within a time series via suffix links and improves AO by explicitly assigning labels to frames linked by suffix links during AO construction.

Given a signal \(O\) sampled at time index \(t\), let \(O[t]\) represent the newly observed value or vector at \(t\). A forward link from state \(i\) to state \(j\), labeled by \(q\), is denoted by \(δ(i, q) = j\). A suffix link from state \(j\) to state \(i\) is denoted by \(sfx[j]/i = j\) without labeling. VMO symbolizes it into a symbolic sequence \(Q = q_1, \ldots, q_t\) with \(T\) states, that denotes the label sequence for labels of observations \(O = O[1]/.../O[T]\). The symbols are formed, by tracking suffix links along the states in an oracle structure (either FO, AO or VMO).

For each observation at time \(t\) of the time series with length \(T\) indexed by a VMO, a suffix link \(sfx[t] = k\) is created pointing back in time \(k\) to where the longest repeated suffix happened. The suffix links also imply a frame-to-frame equivalency between \(j\) and \(k\) given \(sfx[t] = k\) that leads to symbolization of the time-series. In order to track the longest repeated suffix at each time index \(t\), the length of the longest repeated suffix at each state \(t\) is computed by the algorithm described in [22] and is denoted by \(ls[t]\). \(ls[t]\) is essential to the online construction algorithm of an oracle structure and its model selection.

B. Model Selection

A complete reference of the VMO symbolization theory and construction algorithms is found in [41].

Fig. 1 shows an example of the oracle structure obtained with the \(sfx\) method that gives the location of the suffixes, with extreme 0 values and two kinds of links: forward and suffix links. There are two kinds of forward links, internal and external (top oracle in Fig. 1). The internal link is a pointer from state \(t - 1\) to \(t\) labeled by \(q_1\), denoted by \(δ(t-1, q_1) = t\). The function of the forward links is to provide an efficient way to retrieve any of the factors of \(Q\), starting from the beginning of \(Q\) and following a unique path formed by forward links. When the most recent internal forward link is unseen for the previous occurrence of \(q_1\), that is when:

\[
q_{t-1} ≠ q_{t-1} k, \\
q_t = q_{t+k}, \\
δ(t, q_{t+k}) = ∅,
\]

An external forward link \(δ(t, q_{t+k}) = t + k\) is created, which is a pointer from state \(t\) to \(t + k\) (labeled by \(q_{t+k}\), where \(k > 1\)).

The suffix link (bottom oracle in Fig. 1) is a backward pointer that links state \(t\) to \(k\), where \(t > k\), and is denoted by \(sfx[t] = k\). It recognizes repeated patterns in \(Q\), and the condition for when a suffix link is created is that \(sfx[t] = k\) refers to the longest repeated suffix of \(q_1, q_2, \ldots, q_t\) is recognized in \(k\). As can be noticed in the top oracle of Fig. 1, there are no suffix links, because the 0 value being very low, a different symbol is assigned to every frame, hence the condition \(sfx[t] = k\) doesn’t link to any longest repeated suffix.

Conversely in the bottom oracle, there are no external forward links, since \(0\) being very high VMO assigns the same symbol to every frame, so \(q_{t+k} = q_{t+k}\) and no external forward link is created.

With different 0 values, VMO constructs different suffix structures and different symbolized sequences from the signal. To select the one symbolized sequence with the most informative variable-order structure, we use information rate (IR) as the criterion in model selection between different structures generated by different 0 values. IR is an information theoretic measure capable of measuring the information content of a time series in terms of the predictability of its source process on the present observation given past ones. VMO uses the same approach as AO to calculate IR.

![Figure 1](image-url)

Figure 1. Two oracle structures with extreme 0 values: the characters near each forward link represent the assigned labels. In the top oracle structure, the 0 value is 0 or extremely low. In the bottom oracle structure the 0 is very high.
Let $x_1^N = \{x_1, x_2, \ldots, x_N\}$ denote time series $x$ with $N$ observations, where $H(x) = -\sum P(x) \log_2 P(x)$ is the entropy of $x$ and the definition of IR is:

$$IR(x_1^{n-1}, x_n) = H(x_n) - H(x_1^n | x_1^{n-1})$$ (7)

IR is the mutual information between the present and past observations, which is maximized when there is a balance between variation and repetition in the symbolized signal. The value of IR could be approximated by replacing the entropy terms in Equation 1 with a complexity measure $C$ associated with a compression algorithm. This complexity measure is the number of bits used to compress $x_n$ independently using the past observations $x_1^{n-1}$.

Consequently,

$$IR(x_1^{n-1}, x_n) = C(x_n) - C(x_1^n | x_1^{n-1})$$ (8)

Compropr, a lossless compression algorithm based on FO and the length of the longest repeated suffix link (Lrs) is provided elsewhere [19], as is the detailed formulation of combining Compropr, AO, and IR [7]. Fig. 2 visualizes the sum of IR values versus different $\theta$s. A VMO with a higher IR value captures more of the repeating subclips (such as patterns, motifs and themes) than the ones with lower IR values.

C. Recurrence plot from VMO

Given the symbolized sequence $S$ by a VMO, the binary self-similarity matrix (SSM) (VMO-RP) is obtained. The index of a suffix link is a point on the VMO-RP and a repeated sequence is detected as a line since it includes repetitions of length 1, 2, … up to the longest repeated length. This makes VMO effectively find a repetition for variable length non-uniform embedding.

![Figure 2. IR values on vertical axis and $\theta$ on horizontal axis. The solid curve in blue shows the relations between the two measures and the dashed black line indicates the selected $\theta$ by locating the maximal IR value.](image)

![Figure 3. VMO generated RPs. Leftmost ‘applause’ (pleasant), center ‘cheering crowd’ (pleasant), rightmost ‘gun shot’ (unpleasant).](image)

III. CLASSIFICATION OF AFFECT IN AUDIO

Stimuli. A database of various environmental sounds is used to test the proposed method: the International Affective Digitized Sounds (IADS-2) [44]. The IADS-2 is a part of a system for emotional assessment preprocessing and consists of a set of standardized, affective sounds that span a wide range of semantic categories, and are normatively rated on the two-dimensional model of emotions valence and arousal.

Dataset. For each sound, 7 RQA measures were extracted, and they are recurrence rate, determinism, average length of diagonals, longest diagonal excluding the LOI, divergence, entropy of diagonal line lengths and determinism to recurrence rate ratio.

Model of Affect. Affect can be described with basic emotional categories or emotional dimensions. We use Russel's two-dimensional model of valence (V) – arousal (A) described by pleasant – unpleasant for V, and awake – tired for A [46].

Classification Model. In this work we use a feedforward artificial neural net (ANN) [5]. There are no cycles or feedback involved, that is, the output information doesn't travel back to the network. The feedforward ANN is chosen for its simplicity and suitability for the problem studied. The Levenberg-Marquardt and the scale conjugate gradient backpropagation learning algorithms are used, and the validation is performed using the mean squared error function (MSE).

IV. EXPERIMENT

The IADS-2 database consists of 167 non-verbal sounds of 6 seconds duration each at 44100 Hz sampling rate. Human affective labeling on the dimensions of valence and arousal showed that the stimulus evoke reactions across the entire range of each dimension. To ensure consistency in emotional labeling, the affective categories were mapped on the V-A model.

Feature extraction. The CQT of each sound is obtained at 44100 Hz sampling rate, hop length of 512 and 84 bins.

Preprocessing. The dataset is preprocessed before training such that column features are centered to have mean 0 and scaled to have standard deviation 1. No feature selection algorithm is used prior to classification.
Classification. The dataset is divided into 70% training, 15% validation and 15% testing. The classification tasks were implemented using the Neural Network Toolbox in MATLAB, in a multiclass one-versus-all fashion whereby each of the six affective sub-dimensions is in turn considered as positive and negative class. For example, the classification on the arousal dimension consists of 3 tasks: 1) classifying high arousal samples (positive class) versus non-high samples (low and neutral as negative class), 2) low arousal (positive class) versus non-low (high and neutral as negative class), 3) neutral (positive class) versus non-neutral (high and low as the negative class). Once the initial results are obtained, the numbers of neurons and hidden layers as well as the training algorithm are adjusted experimentally to improve the network’s performance. The final results are then averaged to get the classifier's performance for V and A.

Performance evaluation. Over-fitting can be a problem if the dataset is too small or biased. Therefore in addition to the MSE function, the Adaptive Synthetic Sampling (ADASYN) algorithm is applied in order to account for the imbalance in the dataset. The classifier's performance is evaluated before and after dataset rebalancing, using a combination of performance metrics taken from the confusion matrix. These are: accuracy, precision, recall, \( F_1 \)-measure, \( F_2 \)-measure, area under the receiver operating characteristic curve (AUC), as well as Cohen's Kappa. Accuracy (ACC) is a proportion of correct guesses, precision or positive predictive value (PPV) is a measure of exactness and recall or true positive rate (TPR) is a measure of completeness. Accuracy and precision are highly sensitive to data imbalance such that if the class labels of each sample are inverted, i.e., positive samples become negative and vice versa, and a new confusion matrix is built, the results would be poor compared to the original confusion matrix. To cope with this problem four additional measures are computed: Cohen's Kappa (\( \kappa \)) measures how well a classifier performed as compared to how well it would have performed simply by chance, the \( F_1 \)-measure is the harmonic mean of precision and recall, \( F_2 \)-measure sways recall more than precision thus emphasizing the false negative value which is the most critical element of the confusion matrix, and AUC measures the discrimination ability of a classifier for various threshold settings.

V. RESULTS

A. Qualitative description of VMO-RPs

The patterns found in Fig. 3 and 4 provide a description of the temporal behavior of the dynamic trajectories. Fig. 3 shows three plots: the first two for applause and cheering crowd respectively, that correspond to the pleasant emotion ‘happiness’, and the third plot for a gun shot that corresponds to an unpleasant emotion ‘fear’. First we note that the VMO-generated RPs display the best recurrences that give an optimal representation of the signal. Second it can be seen for the gun shot plot, that the unpleasant emotion shows sudden changes in the dynamics of the system, while for the pleasant emotion, the plots show regularities in the patterns. Fig. 4 shows two plots, of a laughing and sobbing man respectively. Again the sudden changes in the dynamic system can be seen for the unpleasant sound, while for the pleasant sound the trajectories show regularities in the patterns.

B. Classification performance

The classifier’s performance was evaluated to compensate for dataset imbalance, with the detailed results and performance metrics on the original dataset are reported in Table 1. The prediction accuracies on (V, A) reported there are (74%, 90%) respectively. In comparison to existing methods, in [47] a classification task was evaluated for the IADS database using a set of 101 acoustic features, achieving a performance of less than 50% accuracy. This shows that our method achieves significant improvement compared to the only existing method we are aware of that is evaluated for the IADS database.

<table>
<thead>
<tr>
<th>Affect</th>
<th>ACC</th>
<th>PPV</th>
<th>TPR</th>
<th>( F_1 )-measure</th>
<th>( F_2 )-measure</th>
<th>( \kappa )</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>0.74</td>
<td>0.93</td>
<td>0.73</td>
<td>0.82</td>
<td>0.76</td>
<td>0.17</td>
<td>0.66</td>
</tr>
<tr>
<td>A</td>
<td>0.90</td>
<td>1.00</td>
<td>0.90</td>
<td>0.94</td>
<td>0.92</td>
<td>0.06</td>
<td>0.72</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this work we have proposed a novel framework for the characterization of affective semantics from the dynamic behavior of auditory scenes. We have extracted the best repeating patterns from symbolized features of the signal. The pattern’s periodicities and complexities were quantified with RQA features derived from the symbolic RP. The features were used in classification achieving a performance of (74%, 90%) on (V, A). For future work, we would like to determine which sounds are better classified with linear and nonlinear features. It may be that sounds that convey affect by changes in time such as regularity in repetition of spectral parts might be more amenable to nonlinear analysis compared to sounds that convey emotions from the timbre of the sound directly. Furthermore it may be interesting to analyze how nonlinear dynamics performs on sounds that include a human voice versus those that are strictly environmental.

REFERENCES


H. He and E. A. Garcia, “Learning from imbalanced data”, IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, 2009, pp. 1263–1284.


