RECURRENCE QUANTIFICATION ANALYSIS FEATURES FOR ENVIRONMENTAL SOUND RECOGNITION

Gerard Roma[1] *Waldo Nogueira*[1 2] *Perfecto Herrera*[1]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain, name.surname@upf.edu
[2] Deutsche Hoerzentrum, Medizinische Hochschule Hannover, Germany

ABSTRACT

This paper tackles the problem of feature aggregation for recognition of auditory scenes in unlabeled audio. We describe a new set of descriptors based on Recurrence Quantification Analysis (RQA), which can be extracted from the similarity matrix of a time series of audio descriptors. We analyze their usefulness for environmental audio recognition combined with traditional feature statistics in the context of the AASP *D-CASE*[1] challenge. Our results show the potential of non-linear time series analysis techniques for dealing with environmental sounds.

1. INTRODUCTION

Technologies for recording digital audio at a reasonable quality have become increasingly available and affordable. As a result, it is now very easy to collect information from our environment through audio recording. Manual segmentation and labeling of audio are labor-intensive tasks. Thus, potential uses of that information are limited by our ability to automatically describe the contents of the recordings.

Research on content-based recognition of environmental sounds has progressed a long way. The problem is usually assigning semantic labels to sound fragments, which makes it easier for humans to interact with recorded audio. The most common approach consists on extracting some features computed from the audio signal, and train a statistical model or machine learning algorithm with some examples for each concept. Most research efforts have proceeded with independent datasets related to diverse applications, which makes it difficult to properly evaluate and compare different methods. In this context, a useful distinction can be made between auditory scenes (i.e. longer recordings of environmental sound) and auditory events (i.e. isolated sounds usually produced by a recognizable source). The AASP *D-CASE* [1] challenge has been proposed for comparing different approaches to both problems. In this article we describe an approach presented to this challenge.

One general problem for classification of audio is feature aggregation. Most common set-ups require the integration of frame-level features over some period of time in order to obtain a single vector that can be input to state-of-the art algorithms for classification and clustering. The typical approach consists on averaging the frame-level features, a process that destroys important information about the temporal evolution and distribution of the features. The development of features that describe the temporal evolution of the sound is still an open issue. In this article we explore the use of Recurrence Quantification Analysis (RQA) [**?**] for supplying some additional information on temporal dynamics.

In the following section we summarize recent work on feature aggregation. We then describe the extraction of RQA features from the evolution of frame-level spectral audio features. The following sections describe our methodology for scene classification and results. We also show some preliminary results on using RQA features for event detection.

2. RELATED WORK

Analysis of environmental audio is a challenging problem as it may potentially include any kind of sounds. Precursors of environmental sound recognition can be found in the literature on sound effects retrieval. Early experiments focused on small datasets assigned to a handful of concepts. Very high classification accuracies were obtained using Hidden Markov Models [3]. However, for larger scales, the most common approach consists of computing means and variances of the frame-level features, often involving one or two derivatives, and using a generic classifier. For example, statistics of a large set of features have been used along with Nearest Neighbour classifiers for large scale applications [4]. One common approach is training the classifier with vector-quantized features [5][6]. A global dictionary is first obtained by clustering a large quantity of frames. Each file is then represented as a normalized histogram of the dictionary items. One problem with this method is that the dictionary is often trained with the whole dataset, before partitioning it into training and test sets. This avoids re-training the dictionary for each fold when an n-fold cross-validation is used for evaluation, but favours overfitting.

The problem of feature aggregation has also been considered in Music Information Retrieval (MIR). An approach to audio temporal integration using Auto-Regressive (AR) models is described in [7]. Unlike AR models, RQA features do not assume linearity or stationarity, while providing model-free, readily usable features. RQA features derived from frame-level chroma features have been tested in the cross-recurrence setting, where two different series are compared, for cover song detection [8]. Our problem is not related to comparison of sequences but to the representation of audio for content-based recognition. A more similar work has tested recurrence time histograms (in this case extracted directly from the audio signal) for genre classification [9].

3. RQA FEATURES

Recurrence Quantification Analysis[2] is a set of techniques developed during the last decade in the study of chaos and complex systems. The basic idea is to quantify patterns that emerge in recurrence plots. RQA has been applied in a wide variety of disciplines. The original technique starts from one-dimensional time series which are assumed to result from a process involving several variables. This multidimensionality is recovered by delaying the

978-1-4799-0972-8/13/$31.00 ©2013IEEE

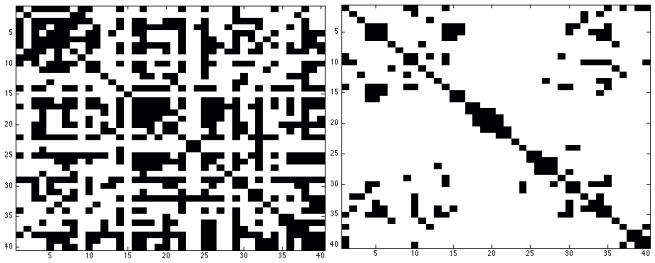

Figure 1: Recurrence plots of *tube* (higher recurrence) and *restaurant* (lower recurrence) auditory scenes.

time series and embedding it in a phase space. The distance matrix of the series is then computed and thresholded to a certain radius r. The radius represents the maximum distance of two observations of the series that will still be considered as belonging to the same state of the system. In the case of audio analysis, it is common to work with multivariate time series such as Mel Frequency Cepstral Coefficients (MFCC). Hence, we adapt the technique by computing and thresholding the similarity matrix obtained from the MFCC representation using cosine distance. Thus, if we denote the series of feature vectors as the multivariate time series X of length N as $X = X_1, X_2, X_3 ... X_N$, then the recurrence plot R is defined as

$$R_{i,j} = \begin{cases} 1 & \text{if } (1 - \frac{X_i \cdot X_j}{||X_i|| ||X_j||}) < r \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Figure 1 shows two of such plots, one corresponding to the *tube* class (with high recurrence levels) and the other to the *restaurant* class (low recurrence but some straight lines). The main intuition is that diagonal lines represent periodicities in the signal, i.e. repeated (or quasi-repeated, depending on the chosen radius) sequences of frames, while vertical lines (or horizontal, since the plot is symmetric) represent stationarities, i.e. the system remains in the same state. From this idea, several metrics have been developed that quantify the amount and length of lines of contiguous points in the matrix. Most features were developed by Ziblut and Webber [10]. We extract the most commonly used ones and add some more variables in order to obtain more features for the classification step.

- Recurrence rate (REC) is just the percentage of points in the recurrence plot.

$$REC = (1/N^2) \sum_{i,j=1}^{N} R_{i,j} \quad (2)$$

- Determinism (DET) is measured as the percentage of points that are in diagonal lines.

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j=1}^{N} R_{i,j}} \quad (3)$$

where $P(l)$ is the histogram of diagonal line lengths l

- Laminarity (LAM) is the percentage of points that form vertical lines.

$$LAM = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)} \quad (4)$$

where $P(v)$ is the histogram of vertical line lengths v

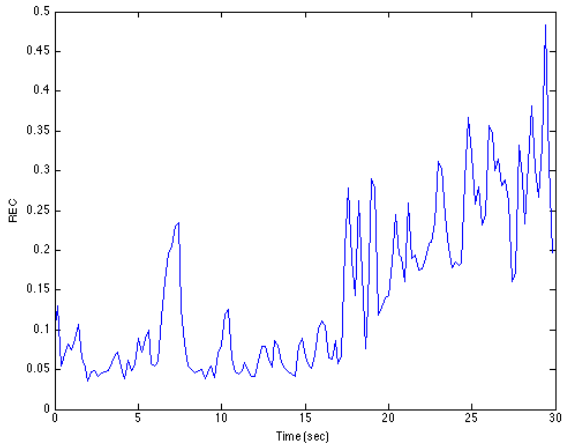

Figure 2: Evolution of REC from a sound of a train departure.

- The ratio between DET and REC is often used. We also use the ratio between LAM and REC, so we define them as

$$DRATIO = N^2 \frac{\sum_{l=l_{min}}^{N} lP(l)}{(\sum_{l=1}^{N} lP(l))^2} \quad (5)$$

$$VRATIO = N^2 \frac{\sum_{v=v_{min}}^{N} vP(v)}{(\sum_{v=1}^{N} vP(v))^2} \quad (6)$$

- LEN and Trapping Time TT are the average diagonal and vertical line lengths

$$LEN = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=l_{min}}^{N} P(l)} \quad (7)$$

$$TT = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)} \quad (8)$$

- Another common feature is the length of the longest diagonal and vertical lines. The inverse of the maximum diagonal (called Divergence) is also used. We use the inverse of both vertical and diagonal maximum lengths

$$DDIV = \frac{1}{max(l)} \quad (9)$$

$$VDIV = \frac{1}{max(v)} \quad (10)$$

- Finally, the Shannon entropy of the diagonal line lengths is commonly used. We also compute the entropy for vertical line lengths.

$$DENT = -\sum_{l=l_{min}}^{N} P(l) ln(P(l)) \quad (11)$$

$$VENT = -\sum_{v=v_{min}}^{N} P(v) ln(P(v)) \quad (12)$$

In order to analyze long series, a windowed version is often used, which consists in computing the recurrence plots from overlapping windows of fix size. This makes it possible to analyze

the temporal evolution of the features. As an example, Figure 2 shows the evolution of REC for a sound of a subway train departing. The first peak corresponds to the door closing signal, and the subsequent increase to the departure. In our experiments, averaging the windowed version as a document-level representation proved to be faster while giving slightly better results. We use texture windows of 400mss, with 50% overlap. With respect to the radius parameter, while it is possible to adjust it taking into account the data (e.g to provide a fixed recurrence rate), we found using a fix value tended to give better classification accuracy. We use a value of $r = 0.03$ (determined experimentally) for the cosine similarity between MFCC vectors. Other parameters are the minimum line lengths (l_{min}, v_{min}). These can be typically set to the minimum 2 points.

4. EXPERIMENTS

RQA features do not require assumptions about linearity or stationarity of the time series. Hence, they seem a good complement for the traditional averaging of spectral audio features in the case of environmental audio. We tested this hypothesis in the context of the *D-CASE* challenge scene classification task. Our experiments are thus aligned with the evaluation setup proposed in the challenge. The development dataset provided by the organizers consist of 100 audio files of 30s divided in 10 auditory scene categories. The task consists in assigning the correct class to unlabelled recordings. In order to further validate our results, we collected a second dataset out of commercial sound effects CDs using the same categories as the challenge database. This in-house database was balanced to 15 examples per class. The system was evaluated using classification accuracy over 5-fold cross-validation using both datasets. A separate test dataset was used the challenge organizers to compare different submissions.

Other than the use of RQA features, our approach is relatively conventional. First, we extract MFCC features from each recording using a sliding window of 25ms with hops of 10ms. MFCCs (13 coefficients from 40 bands) using the rastamat [11] library with a frequency range of 0-900Hz. Extending the frequency range did not improve our results. The resulting frame-level features are averaged to provide the traditional mean and standard deviation features. On the other hand, we obtain RQA features from the same MFCC vectors using a texture window of 400ms. For each window, we obtain a recurrence plot by computing the cosine distance between all MFCC vectors inside the window and removing values above the radius ($r = 0.03$). From this plot we compute the features described in the previous section. We then average the resulting features to obtain a document-level description. Annotation of unseen files is done through an SVM classifier trained on MFCC mean, deviation and RQA features. We use the libSVM [12] implementation with the default RBF kernel. A grid search step is performed for each fold by further partitioning the training data to find optimal values for C and γ.

5. RESULTS AND DISCUSSION

In order to test RQA features for scene recognition, we compared several feature sets feature sets aggregated from the MFCC frames: MV (mean and variance, 26 dimensions), MVD (MV+ delta MFCCs, 39), RQA (11) and $MV + RQA$ (37). Results are shown in Figure 3. Results for the in-house database were expectedly worse, as it was collected from a variety of sources and recording

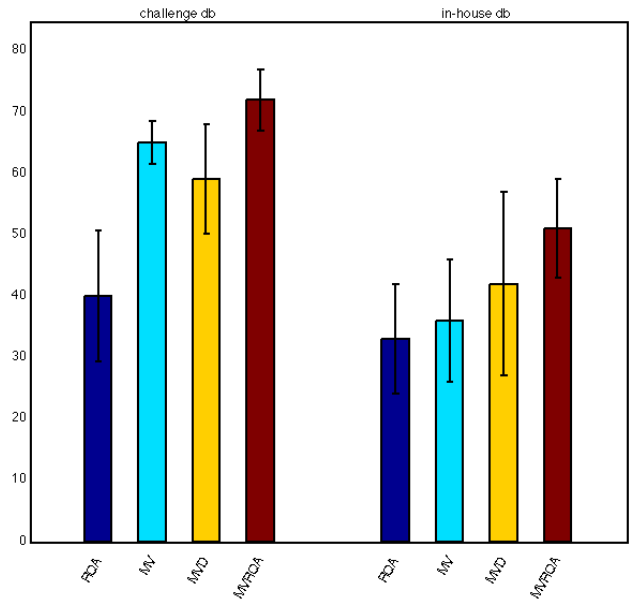

Figure 3: Classification accuracy (mean and standard deviation) of each feature set in the two tested databases

equipments, unlike the challenge database. From our results, it can be inferred that a classifier significantly better than random guessing (10%) can be built with the baseline MV set. Our baseline based on $MFCC + SVM$ also improves on the general $MFCC + GMM$ baseline for the challenge which attained an accuracy of $52 \pm 13\%$. Adding delta features, which are part of many MFCC-based systems does not improve the result for the challenge database but it does for the in-house database. This is possibly related to the differences in durations among databases. Scenes of the challenge database are long (30s) and the rate of change becomes blurred over time. The in-house database includes clips of diverse lengths. Contrastingly, RQA features improve classification accuracy in both cases. This supports the idea that RQA features capture aspects of the short-term evolution of the sound that are not captured by traditional statistics. The system based on $MV + RQA$ was submitted to the challenge, scoring $76 \pm 7.2\%$ on the test dataset. This value is within the standard deviation range obtained with the development dataset. Figure 4 shows the confusion matrix corresponding to a run of the best set with the challenge database. The most common confusions seem quite natural: *tube* is occasionally confused with *tubestation*, and *quietstreet* with *park*. It is obvious that sounds belonging to these categories should be very similar. While we have not conducted formal testing, by listening to the recordings it can be expected that human subjects would also find it difficult to distinguish between some of these classes.

6. EVENT DETECTION

Since our approach involves short windows, we tested it on the event detection task of the challenge. However, due to time restrictions this approach was not presented to the challenge and has not been independently validated. We describe a preliminary experiment using the *Office Synthetic* task[1]. For this task, training files for different event classes were provided. An artificial scene

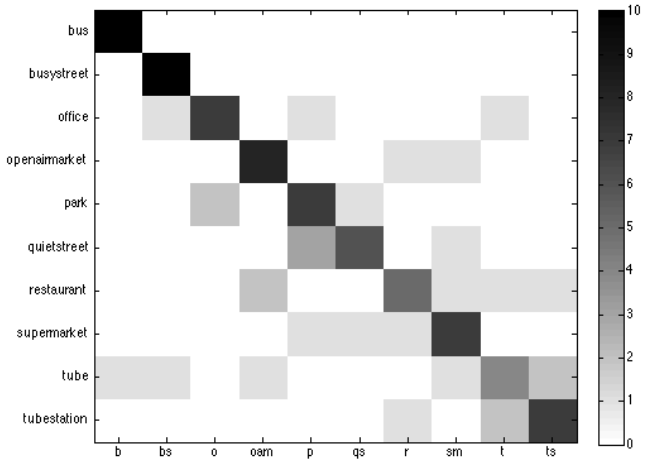

Figure 4: Confusion matrix corresponding to the best feature set

Table 1: Event detection accuracy (MV)

Metric	Frame-Based	Event -based	Class-Based
Rec	0.51	0.39	0.41
Pre	0.49	0.13	0.13
F	0.50	0.20	0.19
AEER	1.49	3.69	3.96

Table 2: Event detection accuracy (MV+RQA)

Metric	Frame-Based	Event -based	Class-Based
Rec	0.52	0.41	0.47
Pre	0.50	0.16	0.17
F	0.52	0.23	0.24
AEER	1.46	3.33	3.41

composed of background and superposed events was then used for testing. Evaluation metrics at the frame, event and class level are precision, recall, F-measure and AEER. For details on the evaluation metrics, we refer to the challenge proposal document [1]. We analyzed the training and test files in a similar way to the scene classification task, but reducing the texture window down to 10 frames (100ms), again with 50% overlap. Increasing the r to 0.05 partially compensated for this loss. Texture windows were classified by the SVM and all frames within the window were assigned to the resulting event class. Tables 2 and 3 show the results of the experiment. An improvement between MV and $MV + RQA$ is observed at a class level but the difference is very small at a frame level. This reflects a compromise between detection of event boundaries and classification accuracy, which in the case of RQA features requires longer time windows. We expect that a better system may be built by having separate paths for boundary detection and event classification.

7. CONCLUSIONS

Recurrence Quantification Analysis is being used in a variety of research areas. The associated features are easily computed from multivariate time series such as spectral audio features. In this article we have shown that RQA features can be combined with conventional feature statistics to increase classification accuracy in the case of auditory scenes. Auditory scenes may involve a wide variety of sounds. By avoiding linearity and stationarity assumptions, RQA features can complement traditional feature statistics and help dealing with the uncertainty. Given their simplicity, we foresee more applications in audio analysis tasks.

8. ACKNOWLEDGEMENTS

This work has been partially funded by the Hearing4all Excellence Initiative.

9. REFERENCES

[1] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Ieee aasp challenge: Detection and classification of acoustic scenes and events," 2013, online web resource. [Online]. Available: http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/

[2] J. Zbilut and C. J. Webber, "Recurrence quantification analysis," in *Wiley Encyclopedia of Biomedical Engineering*, M.Akay, Ed. Hoboken: John Wiley and Sons, 2006.

[3] T. Zhang and C.-C. Kuo, "Classification and retrieval of sound effects in audiovisual data management," in *Signals, Systems, and Computers, 1999. Conference Record of the Thirty-Third Asilomar Conference on*, vol. 1, 1999, pp. 730–734 vol.1.

[4] P. Cano, M. Koppenberger, and N. Wack, "An industrial-strength content-based music recommendation system," in *SIGIR*, 2005, p. 673.

[5] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Multimedia Information Retrieval*, 2008, pp. 105–112.

[6] K. Lee and D. P. W. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 6, pp. 1406–1416, 2010.

[7] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 1, pp. 174–186, 2009.

[8] J. Serrà, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, p. 093017, 09/2009 2009.

[9] J. Serrà, C. A. de los Santos, and R. G. Andrzejak, "Nonlinear audio recurrence analysis with application to genre classification," in *IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, Prague, Czech Republic, 23/05/2011 2011, pp. 169–172.

[10] C. L. Webber and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies," *Journal of Applied Physiology*, vol. 76, no. 2, pp. 965–973, 1994.

[11] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[12] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.