Early frame-based detection of acoustic scenes

Maxime Sangnier, Jérôme Gauthier, Alain Rakotomamonjy

EARLY FRAME-BASED DETECTION OF ACOUSTIC SCENES

Maxime Sangnier[†] * *Jérôme Gauthier*[‡] *Alain Rakotomamonjy*[‖]

[†] LTCI UMR CNRS 5141, Télécom ParisTech
[‡] LADIS, CEA, LIST
[‖] LITIS EA 4108, University of Rouen

ABSTRACT

Let us consider a specific acoustic scene appearing in a continuous audio stream recorded while making a trip a in city. In this work, we aim at detecting at the earliest opportunity the several occurrences of this scene. The objective in early detection is then to build a decision function that is able to go off as soon as possible from the onset of a scene occurrence.This implies making a decision with an incomplete information. This paper proposes a novel framework in this area that i) can guarantee the decision made with a partial observation to be the same as the one with the full observation; ii) incorporates in a non-confusing manner the lack of knowledge about the minimal amount of information needed to make a decision. The proposed detector is based on mapping the temporal sequences to a landmarking space thanks to appropriately designed similarity functions. As a by-product, the built framework benefits from a scalable learning problem. A preliminary experimental study provides compelling results on a soundscape dataset.

Index Terms— Early detection, event detection, scene analysis, machine learning.

1. INTRODUCTION

The problem of recognizing acoustic environments is known as the problem of audio scene classification [1, 2, 3]. This classification task is of primary importance in the domain of machine listening since it is strongly related to the context in which the acquisition device (capturing the audio scene) lives. Typically, in order to get some context awareness, a machine (say a smart-phone or any mobile electronic device) should be able to predict the environment in which it currently resides. In this framework, earliness of the decision is also a major issue so as to improve adaptivity of the machine.

Concretely, early detection is the capability of detecting as soon as possible an occurrence of the soundscape one looks for during an online sequential analysis of an audio stream. This implies making a decision with the incomplete observation of an occurrence (*i.e.* a partial information). Suppose that the acoustic scene looked for is of finite duration, our objective is to build a detector that is able to make a correct decision as soon as an occurrence appears and obviously before it ends. This task is often called "early event detection" in the computer science literature, while "event detection" may have a different definition in the audio and acoustic community.

Even though the general topic of event detection has been explored in several fields like computer vision [4] and disease outbreak [5], *early* event detection has just recently appeared in the machine learning community [6, 7]. This field of early event detection is also strongly related to early classification of time series, which is an active area of research since the early attempts from [8] and latter from [9, 10]. More recently, a framework to classify temporal sequences as soon as possible and with a predefined probabilistic *reliability* has been introduced in [11].

Reliability is the first of two noteworthy concepts for early decision systems. It is defined in a probabilistic way in [11]: *the probability that a prediction based on incomplete information is the same as the one based on the complete information*. This property is essential for early systems since it guarantees the consistency between the decisions with partial and full observations. In the previous studies, reliability is always a starting point but is more or less thoroughly analyzed.

The second important point is *earliness*. The detector is learned so as to make a decision with partial observations but without knowing exactly the sufficient amount of information to collect. A tempting way to achieve earliness is to force partial observations to be well recognized [12, 7]. This is quite computationally demanding and more importantly this is a simplistic way to handle the lack of knowledge about the minimal amount of information required to make a decision. Indeed, such a procedure implies considering partial observations as soundscape occurrences. However, some of those incomplete observations should not be detected since not enough information has been collected yet. This may confuse the learning of the recognition system.

In this work, we describe a novel and general framework to build an early non-linear detector of temporal events, that we apply to acoustic scene analysis. We start with a reliable model, where reliability is considered in a deterministic way: we ensure that the decision with a partial observation is identical to the one achieved with the full sequence. Then, we give a relaxed version of this model. Moreover, we make the hypothesis that the sequences are characterized by unknown discriminative audio frames, where a frame corresponds to a single unit in the temporal sequence. Intuitively, this choice is reasonable: suppose for instance that we want to detect when a person enters a kid play area;then, as soon as an audio frame similar to a cry (or a shout) appears in the audio stream, the early detector can go off.

Such a frame-based approach is well suited to a sequential analysis, since at each time step a new frame is collected. Moreover, it makes a link with a relevant topic in machine learning: Multiple Instance Learning (MIL) [13, 14, 15]. The MIL paradigm is aimed at labelling a bag of instances based on these instances, but without knowing beforehand which instance is discriminative. Thus there are two kinds of labels in MIL: Y, assigned to a bag; and y, which is a latent and unknown label, assigned to an instance. Then, each bag that contains (at least) one instance labelled $y = +1$ (called

*This work has been mainly done when the author was with CEA, LIST. It was partially funded by the *Direction Générale de l'Armement*, French Ministry of Defense. Part of this work has been done while AR was visiting the LIF, Aix-Marseille University.

a witness) gets $Y = +1$. The others are assigned $Y = -1$ (and every instance in them is associated to $y = -1$). There is an ambiguity quite difficult to handle in MIL since only the bags labels Y are known (the latent labels y are unknown). This difficulty is very similar to our framework when we look at a temporal sequence as a bag of frames: determining the minimal amount of information to collect in order to make a decision boils down to discovering the discriminative frames (or the witnesses in MIL). Thus, we leverage this connection in order to propose an early detector that i) non-confusingly handles the lack of knowledge inherent to early detection; ii) does not need to enumerate all partial observations; iii) results in a simple, fast to solve and scalable learning problem.

Besides a short discussion about the closest related works in the forthcoming section, this paper first introduces the proposed framework. We detail how to achieve a reliable and early decision before giving an insight on how to relax the constraints that are too restrictive. The last section is devoted first to a numerical evaluation of this framework and then to a comparison, on a real-world soundscape dataset, to the main approach dealing with early detection in machine learning [7].

2. RELATED WORK

Recently, the problem of early detection has been addressed in [7], resulting in the method called Maximum Margin Early event Detector (MMED). This method is a loss-augmented linear detector which promotes both reliability and earliness. This is achieved by including all partially observed sequences in an extended version of structured output Support Vector Machine (SVM) [16]. In practice, earliness is obtained by penalizing wrong detections for partial sequences. Besides earliness, MMED tries to learn a reliable detector by forcing the decision function to increase as long as it analyzes a scene occurrence. This approximate growth of the decision function conveys the following idea: the more information is collected, the more confident the decision should be. However, the learned detector is not deterministically reliable.

The MIL paradigm is naturally suitable for recognition of temporal sequences. Thus, MIL has been recently applied to early recognition of human actions in a probabilistic setting [12]. Like in [7], the system proposed in [12] tackles earliness through augmented training sequences (partial observations) and does not focus on reliability. On the contrary, our framework does not augment the training dataset and leverage an embedding of the temporal sequences into a landmarking space by means of a non-symmetric non-positive semi-definite (non-PSD) similarity measure with specific properties. This technique, as well as our learning problem, is quite similar to the framework proposed for MIL problems in [17], called Multiple-Instance Learning via Embedded instance Selection (MILES). The main difference between our detector and MILES is the temporality inherent to our bags, which induces modifications of the learning problem in order to handle sequential detection, reliability and earliness.

3. FRAMEWORK FOR EARLY DETECTION

3.1. Problem definition

In this work, the time is discretized and embodied by the subscript t. It can take any value between 1 and T, which is a predefined upper-bound. Each temporal sequence is handled through a discrete time-feature representation $\mathbf{X}_{1..T} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_t}, \ldots, \boldsymbol{x_T})$. This one is a tuple of feature vectors $\boldsymbol{x_t}$ that characterize the sequence at time t. From now on, let $\mathcal{X}$ be a set of feature vectors and time-feature representations.

The aim of this study is to build a real-valued decision function $f\colon \mathcal{X} \to \mathbb{R}$ such that $f(\mathbf{X}_{1..t})$ predicts the nature of the full sequence $\mathbf{X}_{1..T}$, given the partial observation $\mathbf{X}_{1..t}$. Let $b \in \mathbb{R}$ be a detection threshold. $f(\mathbf{X}_{1..t}) \geq b$ claims that the sequence $\mathbf{X}_{1..T}$ is an occurrence of the soundscape we are looking for (this is a detection), while $f(\mathbf{X}_{1..t}) < b$ means that the sequence $\mathbf{X}_{1..T}$ does not represent the acoustic scene or that the detector did not collect enough information to make a decision. This is the default state.

The problem of early detection is to detect a scene occurrence as soon as possible. Concretely, we shall produce a decision $f(\mathbf{X}_{1..t}) \geq b$ with the shortest partial observation $\mathbf{X}_{1..t}$ (the smallest t), only when $\mathbf{X}_{1..T}$ represents the audio scene.

In practice, such a detector is used in a sequential way. A decision is computed at each time step: $f(\mathbf{X}_{1..1})$, $f(\mathbf{X}_{1..2})$, …, $f(\mathbf{X}_{1..t})$. When $f(\mathbf{X}_{1..t}) \geq b$, the analysis is interrupted and a notification of detection is thrown. This means that $\mathbf{X}_{1..T}$ is declared as a scene occurrence if and only if $g(\mathbf{X}_{1..T}) = \max\limits_{1\leq t\leq T} f(\mathbf{X}_{1..t}) \geq b$. Thus, without any assumption, the decision function in the prediction phase g is different from the learned one f . This situation appears for many early detectors [12, 7]. This is so because learning directly g is difficult since it would usually lead to non-convex optimization problems.

The main aim of this paper is to propose a framework where f is *early* and can also be *reliable* (that is f and g produce the same decisions). We suppose that we are provided with a set of training sequences $\left\{\left(\mathbf{X}^{(i)}_{1..T}, Y_i\right)\right\}_{1\leq i\leq n}$, where the label Y_i is equal to $+1$ when the sequence $\mathbf{X}^{(i)}_{1..T}$ is a scene occurrence and to -1 otherwise. Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a similarity measure and $\{\boldsymbol{p_j}\}_{1\leq j\leq m}$ some discriminative frames (called landmarks). Similarly to a kernel machine, in this work, we look for the detector f in $\operatorname{span}\{k(\cdot, \boldsymbol{p_j}), j = 1, \ldots, m\}$, which is the linear span of the evaluation functions $k(\cdot, \boldsymbol{p_j})\colon \mathcal{X} \to \mathbb{R}$ (for $j = 1, \ldots, m$). Thus, f can be defined by $f(\mathbf{X}_{1..t}) = \sum_{j=1}^{m} \mathrm{w}_j k(\mathbf{X}_{1..t}, \boldsymbol{p_j})$, where $\mathbf{w} \in \mathbb{R}^m$ is unknown and has to be learned. When the proximity function k is non-PSD (which will be the case here), the functional f can be learned as a linear function in a landmarking space [18, 19]: $f(\mathbf{X}_{1..t}) = \langle \mathbf{w} \mid \psi(\mathbf{X}_{1..t})\rangle_{\ell_2}$, where $\psi\colon \mathcal{X} \to \mathbb{R}^m$ is a map such that $\psi(\mathbf{X}_{1..t}) = (k(\mathbf{X}_{1..t}, \boldsymbol{p_1}), \ldots, k(\mathbf{X}_{1..t}, \boldsymbol{p_1}))$.

3.2. Reliability

Given the above notation, we now propose a framework to build a reliable detector. Such a detector is characterized by: $[g(\mathbf{X}_{1..T}) \geq b] \Leftrightarrow [f(\mathbf{X}_{1..T}) \geq b]$, where $g(\mathbf{X}_{1..T}) = \max\limits_{1\leq t\leq T} f(\mathbf{X}_{1..t})$. This means that the label output during the sequential prediction phase is the same as the one obtained with the full sequence $\mathbf{X}_{1..T}$.

It is quite interesting to note that such a reliable detector can be easily obtained with few assumptions on the model $f = \langle \mathbf{w} \mid \psi(\cdot)\rangle_{\ell_2}$: suppose that $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a non-decreasing time-dependent similarity measure, that is:

$$\forall \mathbf{X}_{1..T}, \boldsymbol{p} \in \mathcal{X}\colon \big[t_1 \leq t_2\big] \Rightarrow \big[k(\mathbf{X}_{1..t_1}, \boldsymbol{p}) \leq k(\mathbf{X}_{1..t_2}, \boldsymbol{p})\big].$$

If $\mathbf{w} \succcurlyeq 0$, then $f = \langle \mathbf{w} \mid \psi(\cdot)\rangle_{\ell_2}$ is a reliable detector [1].

[1] Indeed, with these assumptions, $\big[t_1 \leq t_2\big] \Rightarrow \big[f(\mathbf{X}_{1..t_1}) \leq f(\mathbf{X}_{1..t_2})\big]$. Thus, $\max\limits_{1\leq t\leq T} f(\mathbf{X}_{1..t}) = f(\mathbf{X}_{1..T})$ and f is reliable.

This statement tells us that by imposing non-negative weights ($\mathbf{w} \succcurlyeq 0$) and by correctly choosing the similarity measure k, the resulting detector f is reliable. The forthcoming paragraph gives a simple recipe to design such an admissible proximity function k.

Consider a frame-to-frame proximity function $q\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, for instance $q(\boldsymbol{x_{t'}}, \boldsymbol{p}) = \langle \boldsymbol{x_{t'}} \mid \boldsymbol{p} \rangle_{\ell_2}$, or $q(\boldsymbol{x_{t'}}, \boldsymbol{p}) = \exp\left(-\gamma \left\| \boldsymbol{x_{t'}} - \boldsymbol{p} \right\|_{\ell_2}^2\right)$, where $\gamma \geq 0$. Then, pool the past proximity values, for example thanks to an ℓ_r-norm: $k(\mathbf{X}_{1..t}, \boldsymbol{p}) = \left(\sum_{t'=1}^{t} \left|q(\boldsymbol{x_{t'}}, \boldsymbol{p})\right|^r\right)^{\frac{1}{r}}$. The resulting similarity measure k is a non-decreasing time-dependent function and thus it can be used to learn a reliable detector. A particular case of this recipe (when q is Gaussian and $r \to +\infty$) gives the function:

$$k(\mathbf{X}_{1..t}, \boldsymbol{p}) = \exp\left(-\gamma \min_{1 \leq t' \leq t} \left\| \boldsymbol{x_{t'}} - \boldsymbol{p} \right\|_{\ell_2}^2\right). \qquad (1)$$

This is a radial basis similarity based on a non-euclidean metric. It has already been used to free oneself from the unknown latent labels in the MIL literature [17].

3.3. Earliness and learning formulation

As already mentioned, since we do not use symmetric positive semi-definite similarity measures, the decision function f is learned in a landmarking space defined by ψ rather than traditionally in the feature space $\mathcal{X}$. This remedy raises a new difficulty: discriminative frames $\{\boldsymbol{p_j}\}_{1 \leq j \leq m}$ are unknown beforehand. A natural way to circumvent this issue [17, 20] is to select the relevant landmarks (during the training) from all the frames available in the training dataset. In practice, this implies initializing the learning procedure with many landmarks $\boldsymbol{p_j}$ and learning a large vector $\mathbf{w}$ with few non-zero components. This can be easily achieved thanks to an ℓ_1-penalization [21].

Another point has not been addressed yet: how to promote earliness of the decision ? A tempting way is to make use of partially-observed sequences in the learning procedure, through an augmented loss function [12, 7]. By forcing them to be well detected, it is possible to control the earliness of the decision. This is a simplistic approach and it usually leads to complex optimization problems which are slow to solve. In the framework presented in this study, a light way to promote earliness is to penalize the selection of latest landmarks while considering only fully observed sequences. These ones are known if, as suggested previously, the landmarks come from the training sequences. Thus, we modify the ℓ_1-penalization to get a weighted regularization $\|\mathbf{w}\|_{\ell_1}^{\boldsymbol{\mu}} = \sum_{j=1}^{m} \mu_j |\mathrm{w}_j|$. Here, $\boldsymbol{\mu}$ is a predefined weighting vector, the components of which are small for early landmarks (typically 1) and progressively larger for later landmarks.

The learning problem of the detector $f = \langle \mathbf{w} \mid \psi(\cdot) \rangle_{\ell_2}$ (jointly with the detection threshold b) is then obtained by writing down an ℓ_1-norm SVM [22] with the features mentioned above (that is, the positivity constraint and the weighted regularization):

$$\begin{array}{ll} \min\limits_{\mathbf{w},\, b,\, \boldsymbol{\xi}} & \|\mathbf{w}\|_{\ell_1}^{\boldsymbol{\mu}} + C \sum_{i=1}^{n} \xi_i \\ \text{s. t.} & \left\{ \begin{array}{l} Y_i \left(\left\langle \mathbf{w} \mid \psi(\mathbf{X}_{1..T}^{(i)}) \right\rangle_{\ell_2} - b \right) \geq 1 - \xi_i, \forall i \\ \boldsymbol{\xi} \succcurlyeq 0,\ \mathbf{w} \succcurlyeq 0, \end{array} \right. \end{array} \qquad (2)$$

where C is a positive tradeoff parameter. Problem (2) is a linear program that can be solved using freely available tools like *lpsolve* [23]. Note that problem (2) defined with the similarity measure (1) has the flavor of the learning problem proposed in [17] but has the extra constraint $\mathbf{w} \succcurlyeq 0$ and the weights $\boldsymbol{\mu}$ in the ℓ_1-penalty.

Despite what has been said before, it is quite important to understand that in some situations, earliness can not be controlled. This is so if the discriminative frames are not expected to appear in a structured manner. For instance if their probability of appearance is equally distributed over the time-frame $[1, T]$. A concrete example is the shout in an audio recording of a kid play area. In this case, we face what we call a non-structured event. Thus, playing with $\boldsymbol{\mu}$ may be harmful and going back to a usual ℓ_1-norm could be in our best interests.

3.4. Empirical relaxation

As it will be revealed through the numerical experiments, reliability is sometimes too restrictive. Even if it is needed theoretically to ensure that the sequential test function g is consistent with the learned one f, reliability may be in contradiction with earliness. Note that the main method we compare our framework to, called MMED [7], is not reliable even though it tries to be.

In a geometrical point of view, the positivity constraint (which induces reliability) forces to select special landmarks. These ones should define a landmarking space in which the occurrences of the event to detect are "above and on the right" of the other sequences. In order to enhance our detector, we propose to replace (when necessary) the constraint $\mathbf{w} \succcurlyeq 0$ in problem (2) by $\frac{1}{m} \sum_{j=1}^{m} \mathrm{w}_j \geq \delta$, where $\delta \geq 0$ is a predefined parameter that controls the *degree of positivity*. In other words, the weights w_j are non-negative "on average". In this way, we build a relaxed version of our detector, which is not theoretically reliable any longer, but which tends to be.

This relaxed model can still be easily learned thanks to the standard decoupling trick: $\|\boldsymbol{\mu}\|_{\ell_1}^{\boldsymbol{\mu}} = \sum_{j=1}^{m} \mu_j ((\mathrm{w}_+)_j + (\mathrm{w}_-)_j)$, with $\mathbf{w} = \mathbf{w}_+ - \mathbf{w}_-$ and $0 \preccurlyeq \mathbf{w}_+, \mathbf{w}_-$. Then we get a new linear program.

4. NUMERICAL EXPERIMENTS

4.1. Evaluation procedure

Several methods are confronted in this section, but always applied to data normalized by their maximum absolute value. The first method is MILES which is a MIL classifier [17]. It takes place in a landmarking space defined by the similarity measure (1). It can be easily extended to a sequential analysis model, even though it has not been designed for this purpose. Then, our approach is called respectively *Sim*X and *Sim*X*R*Y for the reliable and the relaxed versions respectively. *X* is replaced by a number that gives the highest weight μ_j. The lowest is always set to 1. The values in between are defined following a linear trend. For our relaxed approach, *Y* is replaced by the value of the *degree of positivity* δ. In both cases (reliable and relaxed), the similarity measure is still (1), as for MILES. Finally, in order to enrich the numerical comparisons, we also apply MMED both in the feature space (*MMEDF*) and in the landmarking one (*MMEDL*) [2].

[2] MMED is a loss-augmented linear detector, but it can be extended to a non-linear model thanks to an adequate feature mapping, like ψ. In our framework, a frame-based version of MMED (nicknamed *MMEDL*) is built by applying it in our landmarking space: $f(\mathbf{X}_{1..t}) = \langle \mathbf{w} \mid \psi(\mathbf{X}_{1..t}) \rangle_{\ell_2}$. Major differences with our approach still remain: MMED uses an ℓ_2-regularization and asks for partially observed sequences to be relatively well detected. We have used the code provided by the authors.

In the whole section, *AUC* refers to the area under the receiver operating curve obtained for $g(\mathbf{X}_{1..T}) = \max_{1\leq t\leq T} f(\mathbf{X}_{1..t})$ (that is for a sequential test). It measures the overall capability of detection independently of the threshold b (1 notifies a perfect ability). Following [7], we also consider the Activity Monitoring Operating Curve (AMOC). It plots the average Normalized Time to Detect (NTtoD) the scene occurrences versus the False Positive Rate (FPR). This curve is obtained by making the detection threshold b vary. To perform a fair comparison, independently of the tradeoff between accuracy and earliness, we analyze the NTtoD at 0.1 FPR.

Every result presented here (AUC, AMOC curve and NTtoD) is an average on 10 random runs, where the models are evaluated on a test dataset, after being learned on a separate training dataset. The parameter γ that defines the landmarking space is set to 2^{-1}. However, C is obtained through a 5-fold cross-validation (maximizing the AUC) on the following grids: $[2^1, 2^2, \ldots, 2^8]$ for the toy dataset and $[2^3, 2^4, \ldots, 2^{10}]$ for the audio one. For MMED, the values of these grids are multiplied by n since the tradeoff parameter is $\frac{C}{n}$.

Eventually, MMED is trained in accordance with its design, our framework and the other confronted methods: scene occurrences are tagged with the time frame label $[1, T]$ (meaning that the whole sequence is an occurrence) and the other sequences with the time frame label $[0, 0]$ (meaning that there is no occurrence in this sequence). A quick comparison reported a faster training and slightly better results this way compared to drowning the occurrences of the scene in a synthesized audio stream.

4.2. Validation of our approach

This first numerical experiment is aimed at assessing the ability of our detector to promote an early detection. This experiment is performed with a toy dataset, which is made up of two classes. Each class is a linear chirp with an additive Gaussian noise and we use the Mel-Frequency Cepstral Coefficients (MFCCs) computed on a sliding window as time-feature representation. The dataset contains 200 sequences of 20 frames. For each run, half of them are selected for training, and the other half is for testing the models.

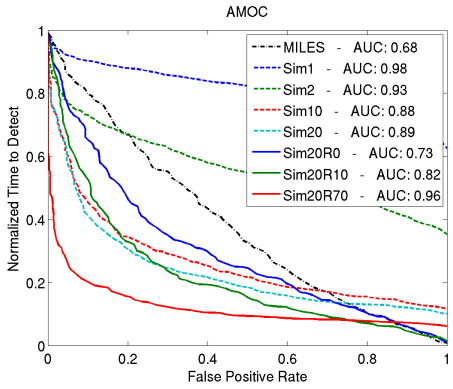

Figure 1: AMOC curve on the toy dataset.

Figure 1 depicts the AMOC curves for MILES and our detector in its reliable (*Sim*) and relaxed (*SimR*) versions. First, let us have a look to the AUC of MILES and *Sim*. Our detector is globally more accurate than MILES. This result validates the role of the reliability property. Moreover, when we penalize more aggressively the later landmarks (from *Sim1* to *Sim20*), the AMOC curve lies down flat (and goes bellow the one of MILES), meaning that the decision is earlier. This gives the green light to our way to promote earliness. On the contrary, the AUC globally decreases (from 0.98 for *Sim1* to 0.89 for *Sim20*), which is an effect of the natural tradeoff between earliness and accuracy. As expected, the relaxed version of our detector (*SimR*) improves detection performances, both in earliness and accuracy.

4.3. Comparison on an audio dataset

Now, we go on with a new experiment: an audio scene detection. This problem, based on the publicly available dataset from [24], consists in deciding in which specific location the audio recordings of 3-s length have been captured. The one-against-all problems we have designed offer 120 sequences for training and 480 for testing, each with 10 frames. The time-feature function used is based on the MFCCs computed on 370-ms overlapping windows [25, section 8.2].

	MILES	Sim2	Sim2R0	Sim2R30	MMEDL	MMEDF
busy street	0.29	0.40	0.24	**0.20**	0.29	0.43
kid play area	**0.16**	0.27	0.19	0.19	0.21	0.31
poolhall	0.22	0.36	0.23	**0.20**	0.52	0.61
restaurant	0.36	0.71	0.31	**0.31**	0.55	0.61
average	0.25	0.44	0.24	**0.22**	0.39	0.49

Table 1: NTtoD at 0.1 FPR for acoustic scene detection.

For this experiment, all the compared methods are as accurate as each others (the average AUC is 0.96 for *Sim10R30*, *MMEDL*, *MMEDS* and 0.95 for MILES, *Sim10R0*). In Table 1, we observe that our reliable detector (*sim2*) is too constrained and achieves a worse earliness than MILES. As already mentioned, this is so because a strong reliability may come as a counterpart of earliness. However, we notice that using our relaxed framework (*SimR*, which only promotes reliability) makes it possible to build the earliest model (*Sim2R30*). Moreover, we remark that, even though MMED is slower than our model to make a decision, it is always earlier when applied in our landmarking space than in the feature space. This can be explained by the frame-based nature of this space.

To conclude this numerical comparison, we shall note that training our early detector (solved directly using *lpsolve* [23]) is as quick as MILES and MMED in the feature space, and dramatically faster than MMED applied in the landmarking space. In both cases, MMED has been trained with Hoai *et al.*'s software, based on CPLEX. This can be explained by the huge number of constraints in MMED, resulting from using partial observations.

5. CONCLUSION

In this paper, we have provided a novel framework for early detection of acoustic scenes. This framework embraces a reliable and a relaxed model, which are both built upon a landmarking space with specific properties. Experimental results highlight that our detector based on similarity functions achieves better performances than its competitor MMED [7].

Constraints relaxations have been considered as a trade-off between theoretical consistency and improved numerical performances. Another way to achieve this goal would be to integrate a representation learning into the global learning strategy. As for us, learning the landmarking mapping would be an important step in the future improvements of this framework.

6. REFERENCES

[1] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[2] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental Sound Recognition With Time-Frequency Audio Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[3] D. Barchiesi, D. Giannoulis, D. Stowell, and M. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[4] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2247–2253, 2007.

[5] D. Neill, A. Moore, and G. Cooper, "A bayesian spatial scan statistic," in *Advances in Neural Information Processing Systems*, 2005.

[6] M. Hoai and F. De la Torre, "Max-margin early event detectors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[7] ——, "Max-margin early event detectors," *International Journal of Computer Vision*, vol. 107, pp. 191–202, 2014.

[8] J. Rodriguez and C. Alonso, "Boosting interval-based literals: Variable length and early classification," in *Workshop on Knowledge Discovery from (Spatio-) Temporal Data*, 2002.

[9] Z. Xing, J. Pei, and P. Yu, "Early prediction on time series: A nearest neighbor approach," in *International Joint Conferences on Artificial Intelligence*, 2009.

[10] ——, "Early classification on time series," *Knowledge and Information Systems*, vol. 31, pp. 105–127, 2012.

[11] N. Parrish, H. Anderson, M. Gupta, and D. Hsiao, "Classifying with confidence from incomplete information," *Journal of Machine Learning Research*, vol. 14, pp. 3561–3589, 2013.

[12] C. Ellis, S. Masood, M. Tappen, J. J. Laviola, and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *International Journal of Computer Vision*, vol. 101, pp. 420–436, 2013.

[13] J. Keeler, D. Rumelhart, and W.-K. Leow, "Integrated Segmentation and Recognition of Hand-Printed Numerals," in *Advances in Neural Information Processing Systems*, 1991.

[14] T. Dietterich, R. Lathrop, and T. Lozano-Prez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, Jan. 1997.

[15] J. Wang and J.-D. Zucker, "Solving the multiple-instance problem: A lazy learning approach," in *International Conference on Machine Learning*, 2000.

[16] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[17] Y. Chen, J. Bi, and J. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1931–1947, 2006.

[18] M.-F. Balcan and A. Blum, "On a theory of learning with similarity functions," in *International Conference on Machine Learning*, 2006.

[19] E. Pekalska and R. Duin, "Beyond traditional kernels: Classification in two dissimilarity-based representation spaces," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, pp. 729–744, 2008.

[20] P. Kar and P. Jain, "Supervised learning with similarity functions," in *Advances in Neural Information Processing Systems*, 2012.

[21] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[22] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm Support Vector Machines," in *Advances in Neural Information Processing Systems*, 2004.

[23] M. Berkelaar, K. Eikland, and P. Notebaert, "lpsolve: Open source (mixed-integer) linear programming system," Eindhoven University of Technology, 2004.

[24] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2015.

[25] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, pp. 4114–4128, 2014.