CHARACTERISATION OF ACOUSTIC SCENES USING A TEMPORALLY-CONSTRAINED SHIFT-INVARIANT MODEL

Emmanouil Benetos[1], Mathieu Lagrange[2], and Simon Dixon[1]*

Centre for Digital Music[1]
Queen Mary University of London
Mile End Road, London E1 4NS, UK
{emmanouilb,simond}@eecs.qmul.ac.uk

Analysis-Synthesis team[2]
IRCAM, CNRS-STMS
1 place Igor Stravinsky, 75004 Paris, France
mathieu.lagrange@ircam.fr

ABSTRACT

In this paper, we propose a method for modeling and classifying acoustic scenes using temporally-constrained shift-invariant probabilistic latent component analysis (SIPLCA). SIPLCA can be used for extracting time-frequency patches from spectrograms in an unsupervised manner. Component-wise hidden Markov models are incorporated to the SIPLCA formulation for enforcing temporal constraints on the activation of each acoustic component. The time-frequency patches are converted to cepstral coefficients in order to provide a compact representation of acoustic events within a scene. Experiments are made using a corpus of train station recordings, classified into 6 scene classes. Results show that the proposed model is able to model salient events within a scene and outperforms the non-negative matrix factorization algorithm for the same task. In addition, it is demonstrated that the use of temporal constraints can lead to improved performance.

1. INTRODUCTION

The problem of modeling acoustic scenes is one of the most challenging tasks in the computational auditory scene analysis (CASA) field [1]. It is closely related to the problem of detecting and classifying acoustic events within a scene, and has numerous applications in audio processing. In the literature the problem is also called *context recognition* [2]. In the case of scene categorisation or characterization, we are interested in specifying the environment of the recording, which is informed by the types of events that are present within the scene of interest. The problem is especially challenging in the case of a real-world scenario with an unlimited set of events which could also overlap in time. It should be noted that event detection and scene categorisation is easily achieved by humans, even in the case of multiple overlapping events.

The literature in this domain is quite vast and we shall now describe two references that consider a technical approach that is close to the one considered in this paper. Mesaros et al. [3] proposed a system for sound event detection which employed probabilistic latent semantic analysis (PLSA) for separating and detecting overlapping events. PLSA (or PLCA, as called in this work) is a factorization technique closely linked to non-negative matrix factorization (NMF). The system was tested in a supervised scenario using a dataset of 103 recordings classified into 10 different scenes, containing events from 61 classes.

In [4], Cotton and Ellis utilised the convolutive NMF algorithm for non-overlapping event detection. A comparison was made between convolutive NMF (which learns spectro-temporal basis matrices) with a frame-based approach using Mel-frequency cepstral coefficients (MFCCs). Experiments performed on a dataset collected under the CHIL project, consisting of 16 different event classes, showed that a combination of the convolutive NMF system and the frame-based system yielded the best results. It should be noted that the convolutive non-negative matrix factorization algorithm is closely related to the shift-invariant probabilistic latent component analysis (SIPLCA) algorithm that is used in the present paper.

In some cases, the salient events that characterise the scene are not known a priori, or may be hard to learn from training data due to the large discrepancy between two acoustic realizations of the same event. For example, in the last decades a wide range of scientific projects designed and put into service massive monitoring devices based on hydrophone or microphone arrays[1]. Among this vast amount of data, one can seek for known acoustic events or alternatively try to discover events of unknown type. The latter leads us to an unsupervised formulation of the scene description problem, where we have only a few loose assumptions about the events of interest and we want the algorithm to be able to extract in an unsupervised manner the events that semantically describe the scene.

Following this approach, Cauchi [5] proposed a method for classifying auditory scenes in an unsupervised manner using sparse non-negative matrix factorization. After extracting spectral basis vectors from acoustic scenes, each basis is converted into MFCCs for compactness. A distance metric is defined for measuring the difference between extracted dictionaries from different scenes. Evaluation is performed on a corpus of 66 recordings taken from several train stations [6], originally created for a perceptual study on acoustic scene categorisation, resulting in six acoustic scene classes. Experiments made by comparing the sparse NMF with a bag-of-features approach from [7] showed that the non-negative matrix factorization algorithm is able to successfully extract salient events within an acoustic scene.

In the present paper, we build upon this work and propose a method for modeling and classifying acoustic scenes in an unsupervised manner using shift-invariant probabilistic models. The shift-invariant probabilistic latent component analysis (SIPLCA)

* The first author is funded by a Westfield Trust research studentship (Queen Mary University of London) and performed part of this work while visiting IRCAM. The second author is partly funded by ANR-11-JS03-005-01. We acknowledge the support of the MIReS project, supported by the European Commission, FP7, ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711.

[1] See for an example: http://www.neptunecanada.ca

algorithm [8] is used in order to extract time-frequency basis matrices from log-frequency spectrograms. In addition, an algorithm is proposed for incorporating temporal constraints to the SIPLCA algorithm using component-wise hidden Markov models (HMMs). These temporal constraints control the occurrence of each acoustic event within a scene using on/off HMMs. The extracted time-frequency basis matrices are afterwards converted to a compact representation using cepstral coefficients. A distance metric is defined for comparing the extracted dictionaries between different acoustic scenes. Evaluation is performed on the same dataset of train station recordings as in [5]. Results using ranking and classification measures show that the proposed SIPLCA models outperform state-of-the art approaches for the same experiment, such as non-negative matrix factorization [5] and a bag-of-frames approach with Gaussian mixture models [7]. In addition, it is shown that incorporating temporal constraints regarding the activation of acoustic scenes, as well as incorporating sparsity constraints on the same activation can lead to more informative basis vectors and thus to improved performance.

The outline of the paper is as follows. The shift-invariant probabilistic latent component analysis method is presented in Section 2. Section 3 presents the proposed temporally-constrained model and the computation of the distance between acoustic scenes. The employed dataset of train station soundscapes, the utilised metrics, and the experimental results compared to other state-of-the-art methods are shown in Section 4. Finally, conclusions are drawn and future directions are indicated in Section 5.

2. SHIFT-INVARIANT PLCA

Shift-invariant probabilistic latent component analysis (SIPLCA) was proposed in [8] for extracting shifted structures from non-negative data. It is a convolutive extension of the probabilistic latent component analysis (PLCA) algorithm, that was proposed by Smaragdis et al. [9]. As PLCA can be viewed as a probabilistic formulation of the non-negative matrix factorization (NMF) algorithm, SIPLCA can be viewed as a probabilistic formulation of the convolutive NMF algorithm [10] using the Kullback-Leibler divergence as a cost function. SIPLCA has been used in the past for pitch tracking [11] and automatic transcription of polyphonic music [12].

The SIPLCA algorithm can support the extraction of a one-dimensional basis from a spectrogram or the extraction of a time-frequency patch. In the present work, we will employ the latter SIPLCA model for extracting 2-dimensional time-frequency patches. The model takes as an input a normalized spectrogram $V_{\omega,t}$ and approximates it as a bivariate distribution $P(\omega, t)$, where $P(\cdot)$ denotes probability, ω is the frequency index and t the time index. $P(\omega, t)$ is decomposed as a series of time-frequency patches convolved over time. The model is formulated as follows:

$$\begin{aligned} V_{\omega,t} \approx P(\omega,t) &= \sum_z P(z)P(\omega,\tau|z) *_\tau P(t|z) \\ &= \sum_z P(z) \sum_\tau P(\omega,\tau|z)P(t-\tau|z) \quad (1) \end{aligned}$$

where $P(\omega, \tau|z)$ is the time-frequency patch for the z-th component, $P(z)$ is the component prior, and $P(t|z)$ is the activation for each component. The unknown model parameters can be estimated using the expectation-maximization algorithm [13]:

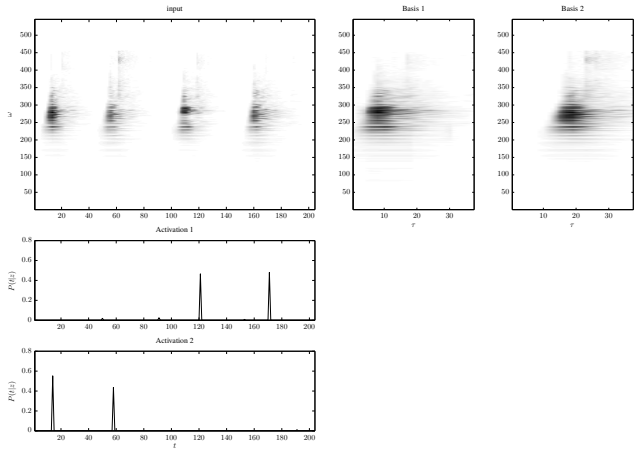

Figure 1: *SIPLCA applied to a sequence of footsteps ($Z = 2$). Top left: input constant-Q transform spectrogram, right: extracted time-frequency patches, bottom: extracted component activations.*

- Expectation step:

$$P(z,\tau|\omega,t) = \frac{P(z)P(\omega,\tau|z)P(t-\tau|z)}{\sum_z \sum_\tau P(z)P(\omega,\tau|z)P(t-\tau|z)} \quad (2)$$

- Maximization step:

$$P(z) = \frac{\sum_{\omega,t,\tau} V_{\omega,t} P(z,\tau|\omega,t)}{\sum_{z,\omega,t,\tau} V_{\omega,t} P(z,\tau|\omega,t)} \quad (3)$$

$$P(\omega,\tau|z) = \frac{\sum_t V_{\omega,t} P(z,\tau|\omega,t)}{\sum_{\omega,\tau,t} V_{\omega,t} P(z,\tau|\omega,t)} \quad (4)$$

$$P(t|z) = \frac{\sum_{\omega,\tau} V_{\omega,t+\tau} P(z,\tau|\omega,t+\tau)}{\sum_{t,\omega,\tau} V_{\omega,t+\tau} P(z,\tau|\omega,t+\tau)} \quad (5)$$

Equation (2) is computed through the model of (1) using Bayes' theorem and expresses the posterior of the unknown variables over the known data. The unknown matrices are initialized with random values. The update rules of (2)-(5) are iterated until convergence. Fig. 1 gives an example of the SIPLCA algorithm, where SIPLCA is applied to a recording of footsteps with $z = 2$. The activations $P(t|z)$ of the footsteps are clearly seen as spikes.

In [14], sparsity constraints are applied to the model in order to provide as meaningful solutions as possible. The sparsity constraints are applied using an *entropic prior*, by modifying the update equations in the maximization step. In the present work, we will encourage sparsity on the component activation $P(t|z)$ in order to derive informative time-frequency patches.

3. PROPOSED METHOD

3.1. Motivation

The motivation behind the model proposed in this paper is to include another level of temporality, which controls the appearance of the time-frequency patches in a recording. These temporal constraints can be supported by incorporating HMMs in the SIPLCA model. Ideally, the component activation function would consist of zeros in case of inactivity and ones at the time instants

where an event would appear. Each HMM can represent a certain component, which would be represented using a two-state, on/off model. This on/off model would serve as an event indicator function, which would enforce temporal constraints in the auditory scene activation matrix. Thus, in this case we propose a novel model which supports time-frequency patches for auditory scene characterization and also controls the temporal succession of events within the scenes.

This work will extend the temporally-constrained convolutive probabilistic model for pitch detection presented in [15], which utilised shift-invariance over log-frequency for spectra instead of performing shift-invariance over time for time-frequency basis as in this work. These models which combine spectral factorization techniques with HMMs were first introduced in [16], where the non-negative HMM algorithm was proposed. Fig. 2 shows the diagram for the proposed system.

3.2. HMM-constrained Shift-invariant PLCA

This proposed temporally-constrained model takes as input a normalized spectrogram $V_{\omega,t}$ and decomposes it as a series of time-frequency patches. Also produced is a component activation matrix, as well as component priors. The activation of a each acoustic component is controlled via a 2-state HMM. The model can be formulated as:

$$V_{\omega,t} \approx P(\omega,t) = \sum_z P(z) \sum_{q_t^{(z)}} P(\omega,\tau|z) *_\tau P(t|z) P(q_t^{(z)}|t) \tag{6}$$

where $q_t^{(z)}$ is the state sequence for the z-th component, $z = 1,\ldots,Z$. Since $\sum_{q_t^{(z)}} P(q_t^{(z)}|t) = 1$, we can revert to the non-temporally constrained model of the previous section. Thus in the model, the desired source activation is given by $P(z|t)P(q_t^{(z)} = 1|t)$.

The activation sequence for each component is constrained using a corresponding HMM, which is based on the produced source activation $P(z,t) = P(z)P(t|z)$. In terms of the activations, the component-wise HMMs can be expressed as:

$$P(\bar{z}) = \sum_{\bar{q}^{(z)}} P(q_1^{(z)}) \prod_t P(q_{t+1}^{(z)}|q_t^{(z)}) \prod_t P_t(z_t|q_t^{(z)}) \tag{7}$$

where $\bar{z}$ refers to the sequence of activations for a given component z, $P(q_1^{(z)})$ is the prior probability, $P(q_{t+1}^{(z)}|q_t^{(z)})$ is the transition matrix for the z-th component, and $P_t(z_t|q_t^{(z)})$ is the observation probability (z_t refers to the activation at frame t). The observation probability for an active component is defined using a sigmoid curve:

$$P_t(z_t|q_t^{(z)} = 1) = \frac{1}{1 + e^{-P(z,t)-\lambda}} \tag{8}$$

where λ is a parameter that controls the component activation (a high value will lead to a low observation probability, leading to an 'off' state). The formulation of the observation function is similar to the one used for multiple note tracking in [17].

As in the model of Section 2, the unknown parameters in the model can be estimated using the expectation maximization algorithm [13]. For the *Expectation* step, we compute the posterior for all the hidden variables:

$$P(z,\tau,q_t^{(1)},\ldots,q_t^{(Z)}|\bar{z},\omega,t) = P(q_t^{(1)},\ldots,q_t^{(Z)}|\bar{z}) P(z,\tau|q_t^{(1)},\ldots,q_t^{(Z)},\omega,t) \tag{9}$$

Since we are utilising independent HMMs, the joint probability for all hidden source states is given by:

$$P_t(q_t^{(1)},\ldots,q_t^{(Z)}|\bar{z}) = \prod_{z=1}^{Z} P_t(q_t^{(z)}|\bar{z}) \tag{10}$$

where

$$P_t(q_t^{(z)}|\bar{z}) = \frac{P_t(\bar{z},q_t^{(z)})}{\sum_{q_t^{(z)}} P_t(\bar{z},q_t^{(z)})} = \frac{\alpha_t(q_t^{(z)})\beta_t(q_t^{(z)})}{\sum_{q_t^{(z)}} \alpha_t(q_t^{(z)})\beta_t(q_t^{(z)})} \tag{11}$$

and $\alpha_t(q_t^{(z)})$, $\beta_t(q_t^{(z)})$ are the forward and backward variables for the z-th HMM [18], which can be computed recursively:

$$\begin{aligned} \alpha_1(q_1) &= P(z_1|q_1)P(q_1) \\ \alpha_{t+1}(q_{t+1}) &= \left(\sum_{q_t} P(q_{t+1}|q_t)\alpha_t(q_t)\right) \cdot P_{t+1}(z_{t+1}|q_{t+1}) \end{aligned} \tag{12}$$

$$\begin{aligned} \beta_T(q_T) &= 1 \\ \beta_t(q_t) &= \sum_{q_{t+1}} \beta_{t+1}(q_{t+1})P(q_{t+1}|q_t)P_{t+1}(z_{t+1}|q_{t+1}) \end{aligned} \tag{13}$$

The second term of (9) can be computed using Bayes' theorem:

$$P(z,\tau|q_t^{(1)},\ldots,q_t^{(Z)},\omega,t) = P(z,\tau|\omega,t) = \frac{P(z)P(\omega,\tau|z)P(t-\tau|z)}{\sum_z \sum_\tau P(z)P(\omega,\tau|z)P(t-\tau|z)|t)} \tag{14}$$

Finally, the posterior for the component transition matrix is given by:

$$P_t(q_t,q_{t+1}|\bar{z}) = \frac{\alpha_t(q_t)P(q_{t+1}|q_t)\beta_{t+1}(q_{t+1})P_t(z_{t+1}|q_{t+1})}{\sum_{q_t,q_{t+1}} \alpha_t(q_t)P(q_{t+1}|q_t)\beta_{t+1}(q_{t+1})P_t(z_{t+1}|q_{t+1})} \tag{15}$$

For the *Maximization* step, the update rules for estimating the unknown parameters are:

$$P(z) = \frac{\sum_{\omega,\tau,t} \overline{\sum}_{q_t^{(z)}} V_{\omega,t} P(z,\tau,q_t^{(1)},\ldots,q_t^{(Z)}|\omega,t)}{\sum_{z,\omega,\tau,t} \overline{\sum}_{q_t^{(z)}} V_{\omega,t} P(z,\tau,q_t^{(1)},\ldots,q_t^{(Z)}|\omega,t)} \tag{16}$$

$$P(\omega,\tau|z) = \frac{\sum_t \overline{\sum}_{q_t^{(z)}} V_{\omega,t} P(z,\tau,q_t^{(1)},\ldots,q_t^{(Z)}|\omega,t)}{\sum_{\omega,\tau,t} \overline{\sum}_{q_t^{(z)}} V_{\omega,t} P(z,\tau,q_t^{(1)},\ldots,q_t^{(Z)}|\omega,t)} \tag{17}$$

$$P(t|z) = \frac{\sum_{\omega,\tau} \overline{\sum}_{q_t^{(z)}} V_{\omega,t+\tau} P(z,\tau,q_t^{(1)},\ldots,q_t^{(Z)}|\omega,t+\tau)}{\sum_{t,\omega,\tau} \overline{\sum}_{q_t^{(z)}} V_{\omega,t+\tau} P(z,\tau,q_t^{(1)},\ldots,q_t^{(Z)}|\omega,t+\tau)} \tag{18}$$

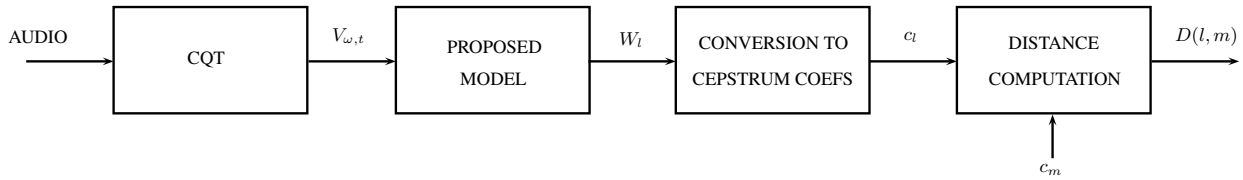

Figure 2: *Diagram for the proposed acoustic scene classification system.*

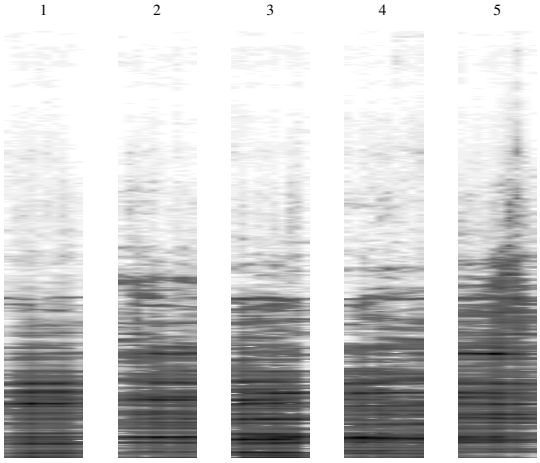

Figure 3: *Extracted time-frequency patches using the proposed model. The recording is part of the dataset by [6], explained in Section 4.*

$$P(q_{t+1}^{(z)}|q_t^{(z)}) = \frac{\sum_t P(q_t^{(z)}, q_{t+1}^{(z)}|\bar{z})}{\sum_{q_{t+1}^{(z)}} \sum_t P(q_t^{(z)}, q_{t+1}^{(z)}|\bar{z})} \tag{19}$$

$$P(q_1^{(z)}) = P_1(q_1^{(z)}|\bar{z}) \tag{20}$$

where $\overline{\sum}_{q_t^{(z)}} = \sum_{q_t^{(1)}} \cdots \sum_{q_t^{(Z)}}$. Eq. (20) updates the component prior using the posterior of eq. (11). Thus, the update equations of the proposed model are a combination of the SIPLCA update rules and the forward-backward HMM algorithm. The final event activation is given by the activation for each component given by the model and the probability for an active state for the corresponding component:

$$P(z, t, q_t^{(z)} = 1) = P(z)P(t|z)P(q_t^{(z)} = 1|t) \tag{21}$$

As in the SIPLCA model of Section 2, sparsity constraints are applied to $P(t|z)$ using the entropic prior of [14] in order to obtain a sparse component activation. In Fig. 3, extracted time-frequency patches can be seen, from a recording employed for evaluation (described in Section 4) using the proposed method with $Z = 5$. Components corresponding to different acoustic events can be seen in the figure. For all the experiments performed in this paper, the length of each basis has been set to 400ms.

3.3. Acoustic Scene Distance

For computing the distance between acoustic scenes, we first compute the constant-Q transform [19] of each 44.1 kHz recording with a log-frequency resolution of 5 bins per octave and an 8-octave span with 27.5 Hz set as the lowest frequency. The step size is set to 40 ms. Afterwards, time-frequency patches are extracted using the proposed HMM-constrained SIPLCA algorithm of Section 3.2 with $Z = \{10, 25, 50\}$ bases and $\lambda = 0.005$ (the value was set after experimentation). Sparsity was enforced to $P(t|z)$ using an entropic prior method of [14] with sparsity parameter values $sH = \{0, 0.1, 0.2, 0.5\}$. In all cases the length of each basis is set to 400 ms.

For each basis $W = P(\omega, \tau|z)$, very small values are replaced by the median value of W. Afterwards, a vector of 13 cepstral coefficients is computed for each basis frame $w[k]$, $(k = 1, \ldots, K)$, in order to result in a compact representation for computational speed purposes. In order to convert a vector $w[k]$ into cepstral coefficients, we employ the formula presented in [20]:

$$c_i = \sum_{k=1}^{K} \log(w[k]) \cos\left(i\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right) \tag{22}$$

where $i = 1, \ldots, 13$. Each vector of cepstral coefficients is then normalized to the range [0,1] region. Thus, the first coefficient that corresponds to the DC component of the signal is dropped. Finally, for each time-frequency basis, the coefficients are summed together over time, thus resulting in a single vector representing a basis. This compressed basis vector is denoted as $W(z)$, where z denotes the component index.

For computing the distance between a scene l and a scene m, we employ the same steps as in [5]. Firstly, we compute the element wise distance between a basis $W_l(z)$, $z = 1, \ldots, Z$ and the nearest basis of dictionary W_m:

$$d_r(l, m) = \min_{j \in [1, Z]} ||W_l(z) - W_m(j)|| \tag{23}$$

The final distance between two acoustic scenes is defined as:

$$D(l, m) = \sum_{z=1}^{Z} d_z(l, m) + d_z(m, l) \tag{24}$$

Equation (24) is formulated in order for the distance measure between two scenes to be symmetric. In the end, the acoustic scene distance matrix D is used for evaluation.

We acknowledge that quantifying the distance between two basis vectors by considering the Euclidean distance of their time average most probably leads to a loss of descriptive power of our model. This choice is made for tractability purposes. Indeed, for the corpus used in this study and 50 bases per item, building the matrix D involves comparing about 10^6 bases. Finding an efficient way of considering the time axis during the distance computation is left for future research.

Scene	Platform	Hall	Corridor	Waiting	Ticket Office	Shop
No. Samples	10	16	12	13	10	5

Table 1: *Class distribution in the employed dataset of acoustic scenes.*

4. EVALUATION

4.1. Dataset

For the acoustic scene classification experiments we employed the dataset created by J. Tardieu [6]. The dataset was originally created for a perceptual study on free- and forced-choice recognition of acoustic scenes by humans. It contains 66 44.1 kHz files recorded in 6 different train stations (Avignon, Bordeaux, Lille Flandres, Nantes, Paris Est, Rennes). Each file is classified into a 'space', which corresponds to the location this file was recorded: platforms, halls, corridors, waiting room, ticket offices, shops. The recordings contain numerous overlapping acoustic events, making even human scene classification a nontrivial task. In Table 1, the class distribution for the employed dataset can be seen. In addition to the ground truth included for each recording, an additional scene label is included as a result of the forced-categorisation perceptual study performed in [6].

4.2. Evaluation metrics

For evaluation, we employed the same set of metrics that were used in [5] for the same experiment, namely the mean average precision (MAP), the 5-precision, and the classification accuracy of a nearest neighbour classifier. The MAP and 5-precision metrics are utilised for ranked retrieval results, where in this case the ranking is given by the values of the distance matrix D. MAP is able to provide a single-figure metric across recall levels and can describe the global behaviour of the system. It is computed using the average precision, which is the average of the precision obtained for the set of top n documents existing after each relevant document is retrieved. The 5-precision is the precision at rank 5, i.e. when the number of relevant samples is equal to 5. It corresponds to the number of samples in the smallest class, which describes the system performance at a local scale.

Regarding the classification accuracy metric, for each row of D we apply the k-nearest neighbour classifier with 11 neighbours, which corresponds to the average number of samples per class.

4.3. Results

Acoustic scene classification experiments were performed using the SIPLCA algorithm of [8] and the proposed SIPLCA algorithm with temporal constraints (TCSIPLCA). Comparative results are also reported using a bag-of-frames (BOF) approach of [7] reported in [5]. The bag-of-frames method computes several audio features which are fed to a Gaussian mixture model classifier. The NMF method of [5] was also implemented and tested. Results are also compared with the human perception experiment reported in [6]. Experiments were performed using different dictionary sizes Z and sparsity parameters sH (details on the range of values can be seen in Section 3.3).

The best results using each employed classifier are presented in Table 2. The proposed temporally-constrained SIPLCA model outperforms all other classifiers using both metrics, apart from the human forced categorisation experiment. The proposed method slightly outperforms the standard SIPLCA algorithm, which in turn outperforms the NMF algorithm. It can also be seen that the BOF method is clearly not suitable for such experiment, since the audio features employed in this method are more appropriate for non-overlapping events, whereas the dataset that is utilised contains concurrent events and a significant level of background noise. However, the human categorisation experiment from [6] outperforms all other approaches.

Model	MAP	5-Precision
Human Perception [6]	0.62	0.73
Random	0.25	0.18
BOF [7]	0.24	0.18
NMF ($Z = 50, sH = 0.99$)	0.32	0.29
SIPLCA ($Z = 25, sH = 0.2$)	0.33	0.35
TCSIPLCA ($Z = 25, sH = 0.2$)	**0.34**	**0.36**

Table 2: *Best MAP and 5-precision results for each model.*

More detailed results for the SIPLCA algorithm using different sparsity parameter values and a different number of extracted bases (Z) can be seen in Fig. 4 (a). In all cases, enforcing sparsity improves performance. It can also be seen that the best performance is reported for $Z = 25$, although the performance of the system using $Z = 50$ improves when greater sparsity on $P(t|z)$ is encouraged. Detailed results for the proposed TCSIPLCA method can be seen in Fig. 4 (b), using different dictionary sizes and sparsity values. It can be seen that the performance reaches a peak when $sH = 0.2$, for the case of $Z = 25$. When using a dictionary size of $Z = 50$, the performance of the proposed method is slightly decreased. Thus, selecting the appropriate number of components is important in the performance of the proposed method, since using too many components will lead to a parts-based representation which in the unsupervised case will lead to non representative dictionaries. Likewise, selecting too few bases will lead to a less descriptive model of the input signal.

Regarding classification accuracy using 11-nearest neighbours, results are shown in Table 3. Again, the TCSIPLCA method outperforms all the other automatic approaches. In this case however, the non-negative matrix factorization approach from [5] outperforms the SIPLCA algorithm by 0.5%. For the TCSIPLCA algorithm, the best performance is again reported for $sH = 0.2$, while for the NMF approach the best performance is reported for $sH = 0$. Regarding dictionary size, the best results are reported for $Z = 50$. Detailed classification results using the SIPLCA and TCSIPLCA methods can be seen in Figures 4 (c) and 4 (d), respectively.

Some experiments were performed by selecting only basis vectors that correspond to a sparse activation $P(t|z)$. In the PLCA domain, the sparseness criterion can be given by maximizing the l_2 norm as in [21], due to the fact that all elements of the activation matrix take values between 0 and 1. However, the performance of the SIPLCA and TCSIPLCA algorithms in fact decreased slightly when selecting only the basis vectors that corresponded to the sparsest activations. This issue may be addressed in the future by enforcing sparsity only to certain components that represent salient events and keeping the rest of the components (which could represent noise) without enforcing sparsity.

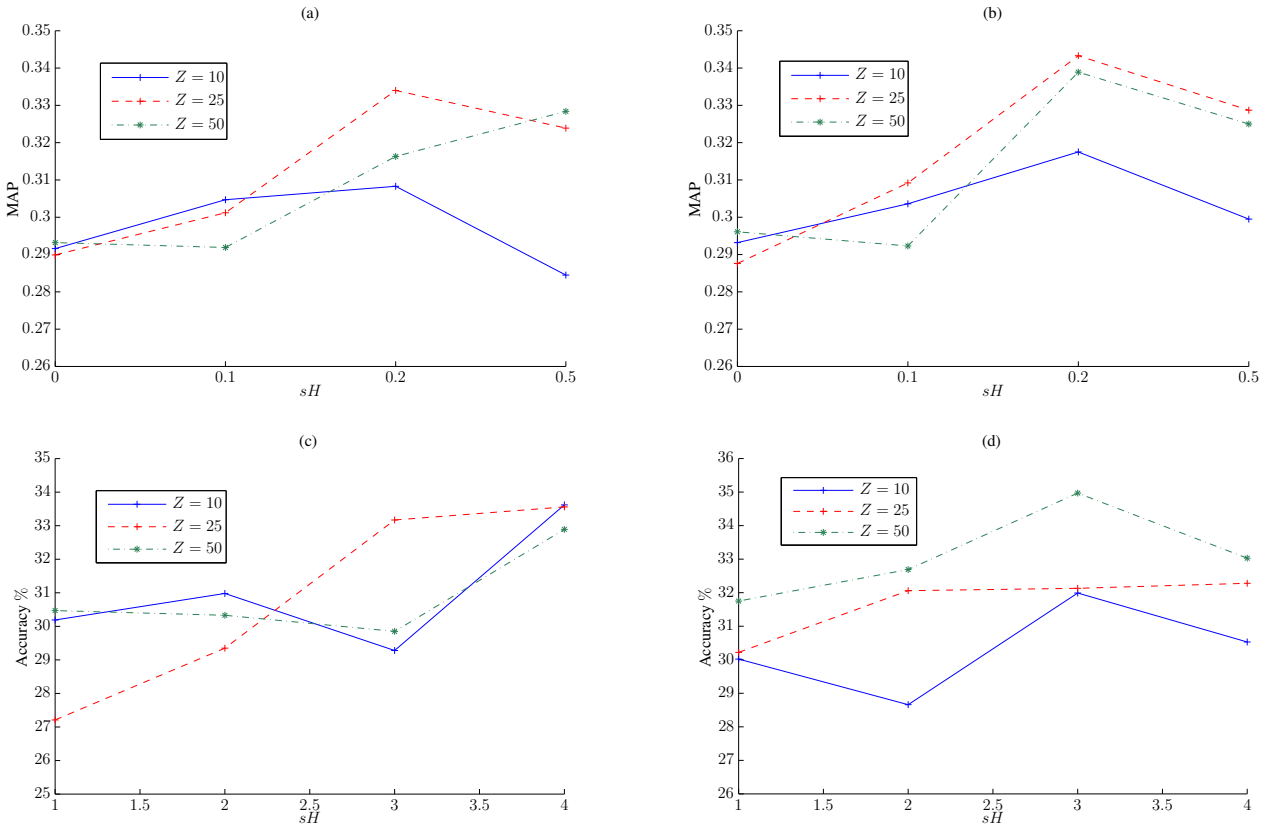

Figure 4: Acoustic scene classification results (MAP) using the (a) SIPLCA and (b) TCSIPLCA algorithm with different sparsity parameters and dictionary size (Z). Classification accuracy (%) using the (c) SIPLCA and (d) TCSIPLCA algorithm with different sparsity parameters and dictionary size (Z).

Classifier	**Accuracy %**
Human Perception [6]	54.8%
Random	16.6%
BOF [7]	19.7%
NMF ($Z = 50, sH = 0$)	34.1%
SIPLCA ($Z = 25, sH = 0.5$)	33.6%
TCSIPLCA ($Z = 50, sH = 0.2$)	35.0%

Table 3: *Best classification accuracy for each model.*

5. CONCLUSIONS

In this work we proposed a method for modeling and classifying acoustic scenes using shift-invariant probabilistic methods. The shift-invariant probabilistic latent component analysis algorithm was utilised for learning time-frequency patches from an input acoustic signal in an unsupervised manner. An algorithm was proposed for incorporating temporal constraints to the SIPLCA model using hidden Markov models, in order to constrain the activation of each event in the signal. In the classification stage, each extracted time-frequency basis is converted into a compact vector of cepstral coefficients for computational speed purposes. The employed dataset consisted of recordings taken from six types of scenes at different train stations. Comparative experiments were performed using a standard non-negative matrix factorization approach, as well as a bag-of-frames algorithm which is based on computing audio features. Results show that using shift-invariant models for learning time-frequency patches improves classification performance. Moreover, incorporating temporal constraints in the SIPLCA model as well as enforcing sparsity constraints in the component activation resulted in improved classification performance.

However, the classification performance of the proposed computational methods is still significantly lower than the human forced categorisation task presented in [6]. We acknowledge that this performance is in our case an upper bound that may not even be reached by purely data-driven methods since humans most probably make extensive use of prior knowledge but the significant gap between the human and computational performances indicates that there is potentially room for improvement on the computational side.

In order to improve spectrogram factorization techniques such as NMF and SIPLCA, additional constraints and knowledge need to be incorporated into the models. A hierarchical model which would consist of event classes and component subclasses would result in a richer model, but would also require prior information

on the shape of each event in order to result in meaningful time-frequency patches. Prior information can be provided by utilising training samples of non-overlapping acoustic events. Also, an additional sparseness constraint could be imposed in the activation matrix, in order to control the number of overlapping components present in the signal (instead of enforcing sparsity as in the present work). In addition, instead of using a first-order Markov model for imposing temporal constraints, a more complex algorithm which would be able to model the duration of each event, such as a semi-Markov model [22] can be employed. Finally, finding an efficient way of comparing extracted time frequency patches is also important. In this respect, we believe that lower bounding approaches to the dynamic time warping technique are of interest [23].

6. REFERENCES

[1] D. L. Wang and G. J. Brown (Eds.), *Computational auditory scene analysis: Principles, algorithms and applications*, IEEE Press/Wiley-Interscience, 2006.

[2] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. 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.

[3] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," in *European Signal Processing Conference*, Barcelona, Spain, Aug. 2011.

[4] C. Cotton and D. Ellis, "Spectral Vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, Oct. 2011, pp. 69–72.

[5] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," M.S. thesis, ATIAM (UPMC / IRCAM / TELECOM ParisTech), Aug. 2011.

[6] J. Tardieu, P. Susini, F. Poisson, P. Lazareff, and S. McAdams, "Perceptual study of soundscapes in train stations," *Applied Acoustics*, vol. 69, no. 12, pp. 1224–1239, 2008.

[7] J.J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[8] P. Smaragdis and B. Raj, "Shift-invariant probabilistic latent component analysis," Tech. Rep., Mitsubishi Electric Research Laboratories, Dec. 2007, TR2007-009.

[9] P. Smaragdis, B. Raj, and Ma. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Neural Information Processing Systems Workshop*, Whistler, Canada, Dec. 2006.

[10] P.D. O'Grady and B.A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, Sept. 2006, pp. 427–432.

[11] G. Mysore and P. Smaragdis, "Relative pitch estimation of multiple instruments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 313–316.

[12] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a convolutive probabilistic model," in *8th Sound and Music Computing Conference*, Padova, Italy, July 2011, pp. 19–24.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[14] P. Smaragdis, "Relative-pitch tracking of multiple arbitary sounds," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3406–3413, May 2009.

[15] E. Benetos and S. Dixon, "A temporally-constrained convolutive probabilistic model for pitch detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, Oct. 2011, pp. 133–136.

[16] G. Mysore, *A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures*, Ph.D. thesis, Standford University, USA, June 2010.

[17] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Advances in Signal Processing*, , no. 8, pp. 154–162, Jan. 2007.

[18] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[19] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference*, Barcelona, Spain, July 2010.

[20] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1933–1941, Mar. 1999.

[21] P. Smaragdis, "Polyphonic pitch tracking by example," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, Oct. 2011, pp. 125–128.

[22] S. Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215 – 243, 2010.

[23] Y. Zhang and J. Glass, "An inner-product lower-bound estimate for dynamic time warping," in *IEEE International Conference on Audio, Speech and Signal Processing*, Prague, Czech Republic, Sept. 2011, pp. 5660–5663.