16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

SINGLE CHANNEL ENUMERATION AND RECOGNITION OF AN UNKNOWN AND TIME-VARYING NUMBER OF SOUND SOURCES

Ilyas Potamitis

Department of Music Technology and Acoustics, Technological Educational Institute of Crete
E. Daskalaki - Perivolia, 74100 Rethymno - Crete, Greece
phone: +30 28310-21911, fax: + 30 28310-23747, email: potamitis@stef.teicrete.gr
web: www.wcl.ee.upatras.gr

ABSTRACT

In this paper we address the problem of on-line enumeration and classification of an audio mix of spectrally overlapping sound sources using a single microphone. We assume that the mix is composed of a subset of realizations of sound sources that belong to an a-priori known set of classes. Each class is represented by a Gaussian mixture model (GMM) probabilistic density function trained from available recordings of each sound class. The approach is based on forming multiple hypotheses on the composition of the mix set that are propagated through time and tested against the likelihood of having produced the audio mix. The likelihood acts as a switch that deletes or reinforces hypotheses. The hypothesis generation and evaluation process is set under a unifying particle filtering framework that estimates the cardinality of the set of sources composing the mix under the maximum a-posteriori (MAP) criterion as well as the sources themselves. The experimental part tests and evaluates the algorithm on real composite environmental soundscenes and on simulations involving rapid changes of cardinality of a set of Gaussian sources.

1. INTRODUCTION

Automatic pattern recognition of general acoustic events is a suitable diagnostic tool for applications involving acoustic surveillance [1], environmental monitoring and biodiversity assessment [2], as well as audio context categorization [3]. However, much of the reported research work is more or less laboratory-based focusing on deriving suitable features [1], [2], and investigating classifiers [1-3], on the problem of classifying sound events that are dominated by sounds belonging to a *single* acoustic class out of a set of classes. The present work reports results towards extending generalized sound recognition to field applications by considering the case of composite sound scenes, that is, simultaneous sound sources that their number and combination may vary in time (e.g., recognition of a dog barking while a bird is singing and a car is passing by).

The automatic recognition of a particular sound event in a soundscene is not a trivial task as the sound sources usually have the same statistical properties and their spectrum may overlap significantly. We consider a soundscene to be an audio mixture that is produced by a process that switches in time between distinct sound events while combining a number of them. The particular sources that constitute the audio mix are thought to be properly gain-scaled realizations out of a-priori known sound classes.

In the case of a single microphone the recognition of the sources becomes even more complicated since:

- There is lack of availability of different versions of the mixture picked-up from many microphones that would allow the construction of an unmixing matrix as in the case of independent component analysis [4] or a receptive beam as in the case of beamforming in microphone arrays [5].
- The number of sources maybe time-varying (e.g. a car is passing by in a soundscene therefore, the number of sources is increased by one and then reduced by one as it leaves, or in a vivid conversation the number of speakers is time-varying and unknown.

The latter complication inflicts many restrictions as most statistical approaches on the problem of source separation assume a known and fixed number of sources composing the audio mix which holds in practice only in specific scenarios.

The purpose of this work is towards building an automatic recognition system of an unknown and time-varying number of general sound events in composite soundscenes (i.e. sound events that overlap in the spectral domain).

The approach is based on having available a set of trained classes of sounds that are supposed to span the audioscene and a search process that locates the combination of classes that best explains the observation of the mixture in terms of the likelihood. The statistical properties of each sound class are represented by a GMM trained on audio corpora of different instances of each sound class. The aim of the GMM is to represent, hopefully, all possible realizations of a particular class and, therefore, the particular realisation taking part in the audio mix. A soundscene is thought to be created by realizations of a subset of the classes. The cardinality of the subset of sources as well as the combination of the sources used to compose the mix is unknown. The cardinality of the sound sources is estimated by first taking the MAP over the posterior pdf of the cardinality. From this optimal cardinality the set of sources that explains better in terms of likelihood the observed mixture is selected. The estimation of the cardinality is based on generating a large number of hypotheses concerning the number and identity of the sources composing the audio mixture and a death/birth/update process for the generation of hypotheses. The whole process is set under the general framework of particle filtering.

2. PROBABILISTIC SINGLE-CHANNEL SOURCE ENUMERATION

We present the mathematical formulation for probabilistic enumeration and recognition of a mixture combination of up to M sources.

2.1 Parametric model of mixture synthesis

Let X_k denote the complex domain of the STFT of the audio mixture and k the frequency-bin index for a fixed-length time window. Let $S_{i,k}$ $i \in [1,..,M]$ be an independent signal source and t is the time-frame index. Then

$$X_k^t = \underbrace{S_{i,k}^t + S_{j,k}^t + .. + S_{n,k}^t}_{M(t) \text{ sources}} \tag{1}$$

where $M(t) \le M$, $\forall t$. One should note that at t+1 sources may appear or disappear thus changing the cardinality of the set of sources composing the mixture as well as the identity of the set of sources that are needed to construct the mix. Even if the cardinality does not change over time the composition of the mixture set under the same cardinality may change (e.g. from [$\mathbf{s}_1$, $\mathbf{s}_3$, $\mathbf{s}_5$] to [$\mathbf{s}_2$, $\mathbf{s}_4$, $\mathbf{s}_5$]).

A common approximation of the power spectrum of the mix can be obtained from (1) by ignoring the cross-terms:

$$\left|X_k^t\right|^2 = \sum_i^{M(t)} \left|S_{i,k}^t\right|^2 \tag{2}$$

Subsequently, a Mel-scale filter-bank is applied to the audio mix observation. The Mel-scale filters apply a linear transformation on the power spectrum by multiplying the power spectral coefficients with positive weights W_k^l [6] and then (2) becomes:

$$\sum_k W_k^l \left|X_k^t\right|^2 = \sum_k W_k^l \sum \left|S_{i,k}^t\right|^2 \tag{3}$$

where, l = 1,2,..,L denotes the filter bank channel and

$$\left|X_l^t\right|^2 = \sum_k W_k^l \left|X_k^t\right|^2,$$

$$\left|S_{i,l}^t\right|^2 = \sum_k W_k^l \left|S_{i,k}^t\right|^2$$

Let $\mathbf{x}$, $\mathbf{s}_i$ be the Mel-scale filterbank power vectors. Then,

$$\mathbf{x}^t = \begin{bmatrix} \left|X_1^t\right|^2 \\ \dots \\ \left|X_L^t\right|^2 \end{bmatrix}, \ \mathbf{s}_i^t = \begin{bmatrix} \left|S_{i,1}^t\right|^2 \\ \dots \\ \left|S_{i,L}^t\right|^2 \end{bmatrix}$$

and (3) becomes $\mathbf{x}^t = \sum_i^{M(t)} \mathbf{s}_i^t$ (4)

In this work we use the Bayesian statistical framework and we incorporate the a-priori information we have for the sources in probability density functions of mixture models for each $\mathbf{s}_i$. Therefore:

$$p(\mathbf{s}_i) = \sum_m w_{i,m} N\left(\mathbf{s}_i; \mu_{i,m}, \Sigma_{i,m}\right),$$

where $\sum_m w_{i,m} = 1$ and the subscripts i, m are indices running over the sources (i = 1,..,M) and the mixtures (m=1,..,m_i) of each source respectively.

Let $\mathbf{S}^t = \left\{\mathbf{s}_1,..,\mathbf{s}_{M(t)}\right\}$ be the set of sources that compose the observation vector at time t according to (4) (e.g. at frame t=10,$\mathbf{S}^{t=10} = \left\{\mathbf{s}_1, \mathbf{s}_4, \mathbf{s}_5\right\}$) and $\mathbf{S}_m^t = \left\{\mathbf{s}_{1,m1},..,\mathbf{s}_{M(t),m_{M(t)}}\right\}$ the set of mixtures of the corresponding set of sources that compose the observation vector at time t (e.g. at frame t=10,$\mathbf{S}_m^t = \left\{\mathbf{s}_{1,12}, \mathbf{s}_{4,6}, \mathbf{s}_{5,2}\right\}$ where the second index is the Gaussian mixture index of the corresponding source). If $\mathbf{S}_m^t$ was known, then from (4):

$$p\left(\mathbf{x}^t \middle| \mathbf{S}_m^t\right) = N\left(\mathbf{x}^t; \sum_i^{M(t)} \mu_{i,m_i}, \sum_i^{M(t)} \Sigma_{i,m_i}\right) \tag{5}$$

However, neither $\mathbf{S}^t$ nor $\mathbf{S}_m^t$ are known. In section 3.2 we describe the process of locating the correct $\mathbf{S}^t$ $\forall t$. It is based on evaluating against (5) all possible combinations of sources.

2.2 Construction of the propagating hypotheses

A power vector observation of a set of a known cardinality $M(t) \le M$, $\forall t$ is thought to be created by first selecting with uniform probability a source $\mathbf{s}_i$ out of M sources and a mixture $N\left(\mathbf{s}_i; \mu_{i,m}, \Sigma_{i,m}\right)$ with probability $w_{i,m}$ and then producing an observation. The same procedure is followed for the rest of the sources up to *M(t)* and the produced observations are added according to (4) to produce the observed audio mix. We have assumed that the subset of sound sources composing the audio mix belongs to a trained set of M models. However, we do not know which and how many of them are combined to create the observed sound mixture at time t. The approach we adopt is to evaluate a number of H hypotheses where H holds all possible combinations of M classes, that is $H=\sum_k M!/(k!(M-k)!)$.

Each hypothesis is a subset of cardinality k. For each source combination we select a mixture according to the prior probabilities of the corresponding source. The indices of the sources and the corresponding selected Gaussian mixtures form a 2-dimensional matrix (see Fig. 1). This matrix constitutes a full hypothesis for the generation of a sound mix and will be set in section 3 under the probabilistic framework of particle filtering. In Fig.1 Hypothesis 1 for example assumes that the observed power mixture is produced by 4 sources. The observation is produced by source 2 - mixture 1, source 5 - mixture 13, source 4 - mixture 32 and source 2 - mixture 9. Hypothesis 2 (note that this hypothesis has different cardinality), assumes that the observation is produced by 2 sources (source 1 - mixture 5, source 3 - mixture 8). There are N such hypothesis that are evaluated in each time-step against the likelihood (5). Hypotheses achieving almost zero likelihood are eliminated while hypotheses scoring high likelihood are reproduced and propagated through time (see section 3). The computation of $\mathbf{S}^t$ is possible since we take all combinations of sources. However, the evaluation of $\mathbf{S}_m^t$ may not be practically possible due to the enormous number of combinations. The approach of constructing $\mathbf{S}^t$ and subsequently evaluating random subsets of $\mathbf{S}_m^t$ achieves its goal due to the fact that Gaussian mixtures of each class represent clusters into the feature space. Therefore if the correct $\mathbf{S}^t$ is located out of all possible combination of sources, this will produce a higher likelihood compared to an incorrect combination of $\mathbf{S}^t$ even if $\mathbf{S}_m^t$ is not exact.

We subsequently describe the selection process of the source and mixtures as well as their rules for the propagation of the hypotheses in time.

At t=1 N hypotheses are generated as follows: $H=\sum_k M!/(k!(M-k)!)$ hypotheses are set to all binomial combinations of the sources. The rest of the hypotheses up to N are repetitions of these combinations. For example if M=4 then the set H, the collection of all finite subsets of the k=4 sources is composed of 15 possible combinations:

H = { [1], [2], [3], [4],	Cardinality 1
[1 2], [1 3], [1 4], [2 3], [2 4], [3 4],	Cardinality 2
[1 2 3], [1 2 4], [1 3 4], [2 3 4],	Cardinality 3
[1 2 3 4]}	Cardinality 4

If N =150 then the initial set of hypotheses (represented in the particle filtering framework in Section 3) will be composed of 10 times H. As they are propagated through time these initial repetitions will become different as their mixtures indices will be resampled (see step (a) below).

For each source of member of H, we select a mixture $N\left(\mathbf{s}_i;\mu_{i,m},\Sigma_{i,m}\right)$ with probability $w_{i,m}$. Therefore, at step t=1 the N generated hypotheses are represented by a 2-D matrix of integer indices as e.g.:

$$\begin{bmatrix} 1 \\ 12 \end{bmatrix}, \begin{bmatrix} 3 \\ 8 \end{bmatrix}, \begin{bmatrix} 2 & 3 \\ 32 & 11 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 3 & 4 \\ 20 & 10 & 6 & 4 \end{bmatrix}$$

Notice that each 2-D matrix holds only indices and not the actual values of the means and variances of the sources.

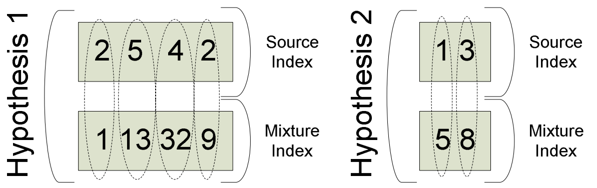

Figure 1 - Illustration of two out of N hypotheses. Each hypothesis is represented by a 2-D matrix of indices. The first row is composed of the source indices proposed to create the mixture and the second row the mixture indices of the corresponding sources responsible for generating the mixture observation.

The actual values of the means and variances of the GMMs of the sources are only needed during the evaluation of the likelihood (5). The true combination of sources that created the observed mixture is definitely included in H because H is composed of all possible combinations of the M sources that are a-priori assumed to span the acoustic space of the audio-scene. At each time step t in the algorithm we randomly select one of the following updates for each hypothesis of the H set:

(a) *Resample the mixtures of all hypotheses*. This step allows exploring the mixture combinations that maximize the likelihood under a fixed source set.

$$\underbrace{\begin{bmatrix} 2 & 3 \\ \mathbf{32} & \mathbf{11} \end{bmatrix}}_{old\ hypothesis} \rightarrow \underbrace{\begin{bmatrix} 2 & 3 \\ \mathbf{12} & \mathbf{11} \end{bmatrix}}_{new\ hypothesis}$$

(b) *Birth of a source in a hypothesis* (only allowed when $M(t)$ < M) with probability pBirth=0.01. A source and a corresponding mixture are added to the current hypothesis. The source to be added is selected with uniform sampling from the set M excluding the sources of the current hypothesis. Let $\boldsymbol{s}_i$ be the selected source. A mixture out of the set of mixtures of $\boldsymbol{s}_i$ is selected with probability $\mathrm{w}_{i,m}$. For example:

$$\underbrace{\begin{bmatrix} 2 & 3 \\ 32 & 11 \end{bmatrix}}_{old\ hypothesis} \rightarrow \underbrace{\begin{bmatrix} 2 & 3 & \mathbf{1} \\ 32 & 11 & \mathbf{8} \end{bmatrix}}_{new\ hypothesis}$$

(c) *Death of a source in a hypothesis* (only allowed when $M(t)$>0) with probability pDelete=0.01. A source and its corresponding mixture are deleted from the current hypothesis. The source to be deleted is selected with uniform sampling out of the set of the previous hypothesis. For example:

$$\underbrace{\begin{bmatrix} 1 & 2 & \mathbf{3} & 4 \\ 20 & 10 & \mathbf{6} & 4 \end{bmatrix}}_{old\ hypothesis} \rightarrow \underbrace{\begin{bmatrix} 1 & 2 & 4 \\ 20 & 10 & 4 \end{bmatrix}}_{new\ hypothesis}$$

(d) *Update of a hypothesis* with probability pUpdate=0.98. The set of sources remains the same. The probability of this step denotes that most of the time the cardinality is expected to hold a fixed value in a soundscene corresponding to the steady state of an acoustic event. Steps (b) and (c) focus on the transition from one steady state to another steady state with probably different cardinality.

(e) *Injection of set H* in the set of hypothesis at random positions selected uniformly out of N. Since it is known that the best combination is inside the set of combinations the initial combinations are re-injected in the pool of hypotheses because some of them may have been eliminated through time. This step is helpful in the case of a jump in cardinality after a long time of stability.

(f) *Calculation of the likelihood of all hypotheses* according to (5) and resampling of the hypothesis set according to their likelihood score. Hypotheses are deleted or duplicated according to the likelihood they score (see section 3).

Subsequently, the time step is increased by one and the stages (b)-(f) are repeated until the end of all observations.

3. HYPOTHESES FILTERING & RECOGNITION

The previous description can be actually set under the umbrella of particle filtering.

For our problem a particle represents the cardinality of the set of sources and i=1 to N is the particle index.

$$c_t^{\mathrm{i}} \sim p\left(k_t \middle| \mathbf{x}^t\right).$$

Let as assume that we are able to simulate N independent particles $\left\{c_{\mathrm{t}}^{\mathrm{i}}\right\}$ $i=1,..,N$ from $p\left(c_t \middle| \mathbf{x}_t\right)$. Then if we find the maximum of the mode of $p\left(c_t \middle| \mathbf{x}_t\right)$ we can estimate the cardinality of the set using the MAP criterion, that is,

$$\hat{c}_t = \arg\max p\left(c_t \middle| \mathbf{x}^t\right).$$

The initial set of sources H at t=0 is composed of all sources combinations. We introduce a drift in the cardinality of all hypotheses at each time-step as:

$$c_t^{\mathrm{i}} = c_{t-1}^{\mathrm{i}} + \Delta c \text{ for i=1,..,}N$$

where

$$\Delta c = \begin{Bmatrix} \text{-1} & \text{death of source, pDeath=0.01} \\ 0 & \text{update of cardinalty, pUpd=0.98} \\ +1 & \text{birth of source, pBirth=0.01} \end{Bmatrix}$$

Using the Birth/Death/Update of sources according to section 3 we construct an hypothesis on set of sources $\mathbf{S}_{i,m}^t$ (see section 2.1 for the definition of $\mathbf{S}_m^t$) for each particle i=1,..,N. Note that $\mathbf{S}_{i,m}^t$ evolved from $\mathbf{S}_{i,m}^{t-1}$.

Particle filtering approximates the density with a sum of Dirac functions centred on the particles.

For i=1,..,N we evaluate the different hypotheses by applying a weight on them given by (5), that is:

$$w_i^t \propto p\left(\mathbf{x}^t \middle| \mathbf{S}_m^t\right)$$

that is, the weight is proportional to the likelihood function and we subsequently normalize it so that $w_i^t = \dfrac{w_i^t}{\sum_j w_j^t}$

Finally, we resample with replacement N samples according to the importance weights, therefore,

$$p\left(c_t \middle| \mathbf{x}^t\right) = \sum w_i^t \left\{\mathbf{S}_{i,m}^t\right\}$$

Once we have acquired the MAP estimate of the cardinality we investigate which sources compose the sets that achieve the optimal cardinality and therefore achieve in parallel the recognition of the sources. This is performed by taking the histogram of all sources from all hypotheses having the MAP cardinality and selecting the first $\mathrm{k_{MAP}}$ sources having the maximum number of occurrences.

The algorithm is real-time even for a reasonably large number of particles since only the Gaussian mixture indices of the sources are propagated through time and not the actual means and variances. The means and variances are only calculated once per cycle of the algorithm in (5).

4. NUMERICAL STUDIES AND EXPERIMENTS

In this section we study the recognition performance of the algorithm using both simulated and true data. Due to the unknown cardinality of the set of sources that produces the observed audio mixture there can be two sources of error. There can be errors in the estimated cardinality (e.g. the true set composing the mixture is [2 3 1] and the estimated set is [1 2 3 4]) and errors in the estimated composition of the set (e.g. sets [1 4 5] and [1 2 5] have the same cardinality but different composition). In order to capture both sources of error we follow [7] and we establish two measures of recognition accuracy: The accuracy of the recognized cardinality $p\left(c_{1:t} = \hat{c}_{1:t}\right)$ and the conditional recognition accuracy of the source composition given the correct cases of cardinality of the sources. In order to have a full picture of the system one should view both figures of error patterns.

4.1 Simulation experiments

In Table 1-2 we present simulation results for the recognized cardinality and composition of sources. The simulation is based on 100 Monte Carlo runs involving mixtures of up to 5 sources. Each source is a 1-D Gaussian mixture of 8 components having random means, weights and variances. The correct cardinality undergoes jumps every 10 time steps. Each experiment is based on 20 sections of 10 time steps (200 time-steps). One should not that this would be an extreme scenario for real audio mixtures as it would correspond to changes in the composition of the mix every 10 frames that is, in a fraction of every second. In Figure 2 – *top* we present a typical single run and in Figure 2 – *bottom* the average error in each time step over 100 Monte Carlo runs. This figure demonstrates that an error is more probable to appear as the mixture undergoes a change of cardinality. This is to be expected as most of the particles during a time period with a fixed cardinality will represent the same cardinality and in an abrupt jump the new cardinality will lead to elimination of most particles. Therefore, the most probable position of an error is during a cardinality jump. However the particle filter quickly recovers and the errors in cardinality are isolated with no consistent loss of track.

# particles	*pUpdate=0.98*		*No birth/death*	
	Mean	*Std*	*Mean*	*Std*
50	77.41	0.05	77.57	0.05
100	91.16	0.03	91.01	0.04
150	92.20	0.03	91.44	0.03
1500	92.20	0.03	91.42	0.04

Table 1. Mean and standard deviation results on the estimated cardinality over 100 Monte Carlo runs for the 5 sources case. We investigated the cases of different number of particles as well as the case of having no birth/death moves (probUpdate=1).

# particles	*pUpdate=0.98*		*No birth/death*	
	Mean	*Std*	*Mean*	*Std*
50	94.74	0.02	94.34	0.02
100	98.12	0.02	98.29	0.01
150	98.29	0.01	98.41	0.01
1500	98.44	0.02	98.51	0.01

Table 2. Mean and standard deviation results on the recognized sources given the correct cardinality over 100 Monte Carlo runs.

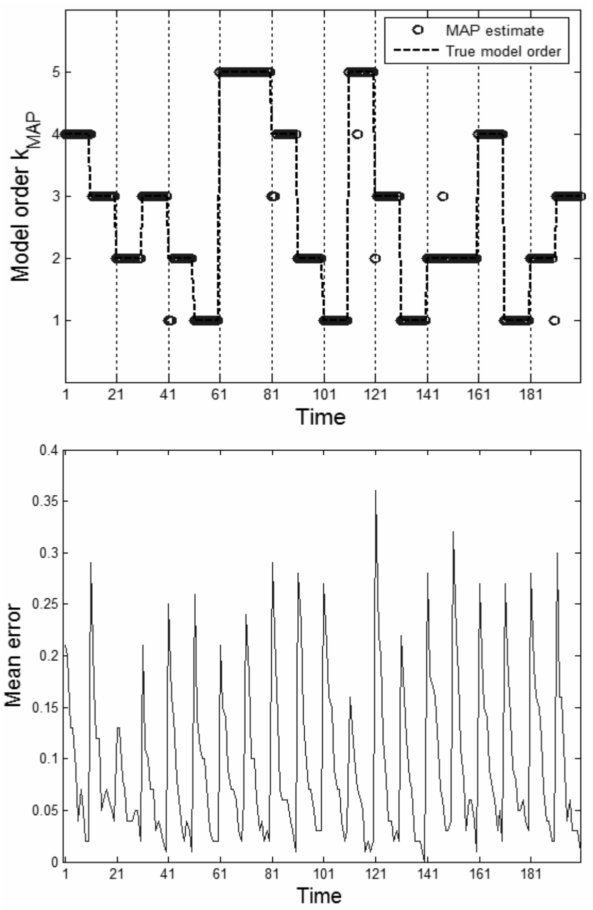

Figure 2 – *Top figure:* Tracking the cardinality of the mixture set over time. The mixture set is composed of mixtures of up to 5 sources that undergoes a random cardinality jump every 10 time steps.

Bottom figure: Mean error over 100 Monte Carlo runs. Most of the errors occur when the cardinality undergoes a jump. The error significantly drops in the steady state.

4.2 Real audio sources experiment

In Fig. 3 we performed an experiment with real audio data sampled at 8000 kHz. An audio mixture is constructed with a dog barking while a bird is singing. The classes of sounds are 4: barking dogs, singing birds; cicadas (an insect species) and rain recordings. 10 recordings of 15 seconds each are taken from the BBC Sound Effects Library. The dimensionality of the STFT is reduced through a filterbank (from 256 to 23 if we apply a 512 samples FFT and a filterbank of 23 bands) allowing the training of more efficient models. The overlap of the window of the short-time Fourier transform is 50%. The mixture model is composed of 16 diagonal covariance mixtures. Figure 3 is produced by using 1000 particles. The models are initialized using 5 iterations of the K-means algorithm and trained using a standard version of the expectation-maximization algorithm.

The segmentation results are very promising as the correct order of the mixture is predicted correctly most of the times (see Fig. 3). There are no algorithms, at least to the knowledge of the author that can be applied to a similar task so as to perform comparative experiments.

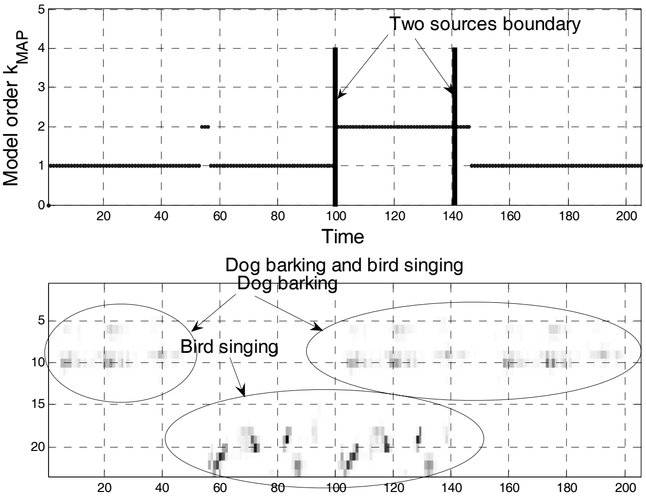

Figure 3 – A dog is barking until frame 55 (manual segmentation). A bird is singing in frames 55-100. From frame 100-142 a dog is barking and a bird is singing. From frame 142-205 a dog is barking. *Top:* Predicted cardinality of the audio event. *Bottom:* Spectrogram of the sound mixture.

5. CONCLUSIONS

In this paper we address the problem of on-line enumeration and classification of the sound sources composing complex sound mixtures using a single microphone. We assume that the audio mix is produced by a subset of realizations of sound sources that belong to a known set of classes. Each class is represented by a Gaussian mixture model (GMM) probabilistic density function trained from available recordings of each sound class separately. The approach is based on forming multiple hypotheses on the cardinality and composition of the set of sound sources that is propagated through time and tested against the likelihood of having produced the power spectrum of the audio mix.

ACKNOWLEDGMENT

This work was supported in part by the E.C under the 7th framework grant Prometheus 214901.

We acknowledge the use of parts of code kindly made available by Prof. Nando de Freitas http://www.cs.ubc.ca/~nando/

REFERENCES

[1] Cowling, M., and Sitte, R., “Comparison of techniques for environmental sound recognition”, *Pattern Recognition Letters*, pp. 2895-2907, vol. 24, no. 15, 2003.

[2] Riede, K., “Acoustic monitoring of Orthoptera and its potential for conservation”, *Journal of Insect Conserv.*, 2 (4), pp. 217–223, 1998.

[3] Eronen, A., et al., “Audio-Based Context Recognition”, *IEEE Trans. on Audio, Speech, and Lang. Proc.*, pp. 321-329, vol. 14, no. 1, 2006.

[4] Hyvärinen, J. Karhunen, E. Oja, “Independent Component Analysis”, *John Wiley & Sons*, 2001.

[5] Michael Brandstein, Darren Ward (editors), “Microphone Arrays: Signal Processing Techniques and Applications”, *Springer*, 2005.

[6] Deng, L., Droppo, J., and Acero, A., “Estimating Cepstrum of Speech Under the Presence of Noise Using a Joint Prior of Static and Dynamic Features”, *IEEE Trans. on Speech & Audio Proc.*, pp. 218-233, vol. 12, no. 3, 2004.

[7] Ba-Ngu Vo; Wing-Kin Ma; Singh, S., “Localizing an unknown time-varying number of speakers: a Bayesian random finite set approach”, *IEEE ICASSP '05*, vol.4, pp. 1073-1076 Vol. 4, 18-23, March 2005.