International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:5, No:11, 2011

Automatic Recognition of an Unknown and Time-Varying Number of Simultaneous Environmental Sound Sources

S. Ntalampiras, I. Potamitis, N. Fakotakis, and S. Kouzoupis

***Abstract*—**The present work faces the problem of automatic enumeration and recognition of an unknown and time-varying number of environmental sound sources while using a single microphone. The assumption that is made is that the sound recorded is a realization of sound sources belonging to a group of audio classes which is known a-priori. We describe two variations of the same principle which is to calculate the distance between the current unknown audio frame and all possible combinations of the classes that are assumed to span the soundscene. We concentrate on categorizing environmental sound sources, such as birds, insects etc. in the task of monitoring the biodiversity of a specific habitat.

***Keywords*—**automatic recognition of multiple sound sources, enumeration of sound sources, computational ecology.

I. Introduction

THE technology of generalized sound recognition as a branch of computational auditory scene analysis [1] can offer reliable solutions to a wide range of applications such as acoustic surveillance [2], context recognition [3], as well as memory aid [4]. Lately, the particular scientific domain has gained the attention of many researchers while the two main problems that need special care are:
a) as the number of the sound classes increases, the recognition performance rapidly decreases and,
b) composite real-life soundscenes can be very difficult to analyze due to the unbounded and time varying number of co-existing audio classes. Most of the previous work in the area is more or less laboratory-based and focus on classifying a single dominating audio source that belongs to a fixed, predefined set of audio classes.

In this work we report results towards expanding sound recognition to field applications which consist of simultaneous, spectrally overlapping sound sources that their number and combination may vary in time (e.g. a bird is singing while rain is falling and a wind is present). We try our approach on a biodiversity monitoring task that involves sound sources encountered in nature such as animal vocalizations, insects, rain and wind sounds etc. Acoustic monitoring can be used to provide baseline information about specific groups of acoustically active biota and automatically construct an inventory of species taxa based on their vocalizations.

The task of enumeration and recognition of audio sources as described above can be characterized by a high degree of difficulty since it is usual for the sound sources to have similar statistical properties while their spectral content usually overlaps significantly. With respect to our work a soundscene is considered to be an audio mixture which is a result of a process that switches in time between distinct sound sources while a random combination of them is selected. We assume that these sources belong to a-priori known set of classes. In our classification approach, the number of sound sources as well as their identities are allowed to vary in time. Our approach is based on having an inventory of audio classes that are assumed to span the acoustic scene and testing each unknown recorded audio segment against all possible combinations of the audio classes. We present two variations that differ in how an audio class is represented and how the distance is calculated (see also Fig. 1):
a) in the first approach we derive the spectral signatures of the set of known classes. Each class is simply represented by the mean power spectrum of available recordings. All possible combinations are produced by adding the corresponding mean power spectrums. A simple distance metric is used to assign unknown frames to classes.
b) The second approach is an elaboration of [5] and requires the construction of GMMs fitted to the power spectrum of each class in order to represent the audio classes that are assumed to span the soundscene. As all combinations are additive in the power domain the likelihood of every possible combination having realized the unknown recording is shown that can be expressed in closed form.

The rest of the paper is organized as follows: Section 2 analyses the two different methods that have been employed on the specific task while Section 3 presents our experiments. Finally our conclusions are drawn in the last Section of the present paper.

II. Analysis of the Recognition Methodology

This section presents the mathematical formulation as regards to the problem of enumeration and recognition of an audio mixture composed of M audio sources.

S. Ntalampiras is with the Electrical and Computer Engineering Department of the University of Patras (phone: +30 2610 996496; fax: +30 2610 997336; e-mail: sntalampiras@upatras.gr).
I. Potamitis is with the Department of Music Technology and Acoustics of the Technological Educational Institute of Crete (phone: +30 28310 21911; fax: +30 2810 58323; e-mail: potamitis@wcl.ee.upatras.gr).
N. Fakotakis is with the Electrical and Computer Engineering Department of the University of Patras (phone: +30 2610 996496; fax: +30 2610 997336; e-mail: fakotaki@upatras.gr).
S. Kouzoupis is with the Department of Music Technology and Acoustics of the Technological Educational Institute of Crete (phone: +30 28310 21911; fax: +30 2810 58323; e-mail: skouzo@staff.teicrete.gr).

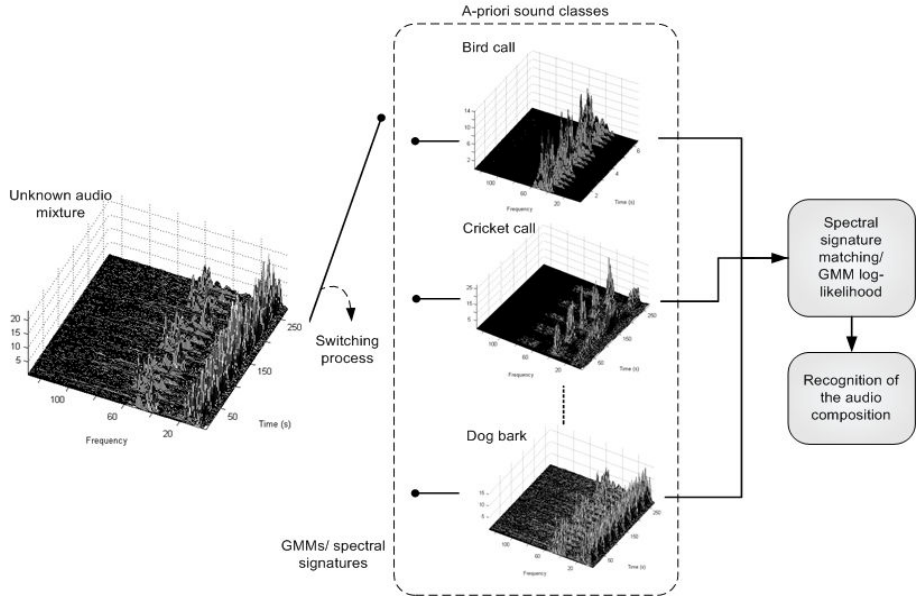

Fig. 1 The unknown recording is a realization of a combination of a set of prototypes. All combinations of audio prototypes are tested of having produced the unknown realization. The combination that produces the lowest distance (method A) and the highest likelihood (method B) is selected

A. Modeling of the Audio Mixture

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 $S_{i,k}$, where $i \in [1,..,M]$ be an independent signal source and t is the time-frame index. Then

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

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 mixture. 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:

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

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:

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

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

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

B. Modeling the known audio classes with the mean spectrum

The general principle of content-based sound recognition is based on the fact that a sound source emits consistent acoustic patterns with a very distinctive and characteristic way to distribute its energy over time on its composing frequencies.

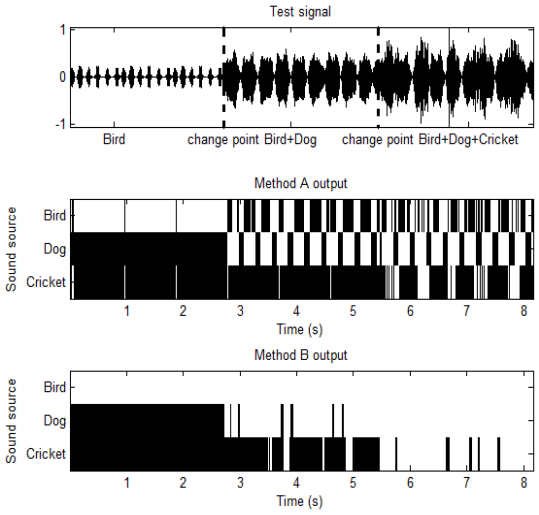

Fig. 2 T*op*: A test signal that progressively increases its number of sources from 1 to 3 after a marked change-point (bird, bird+dog, bird+dog+cricket). *Middle* and *Bottom*: The response of *Methods A and B* are depicted respectively where white spaces correspond to active sound sources

This constitutes the so-called *spectral signature* of the specific sound source and can be used as the fingerprint of a specific source [7-10]. The first approach models the spectral signature of each sound class by using the mean-power of all available recordings of each audio class. Subsequently, the Mel-scale spectrum of each sound class is derived according to (3). The filterbank is essential for reducing the variability of the spectrum of each source. The next step is to derive the spectral signature for every possible combination in the power domain according to (4). For example if [$\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$] are the a-priori classes assumed to span the audio scene the possible combinations are [$\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$, $\mathbf{s}_1$+$\mathbf{s}_2$, $\mathbf{s}_1$+$\mathbf{s}_3$, $\mathbf{s}_2$+$\mathbf{s}_3$, $\mathbf{s}_1$+$\mathbf{s}_2$+$\mathbf{s}_3$]. Unknown content can then be identified by comparing its signature to the signature of each member of the set of combinations. The distance metric essentially is the absolute difference between the unknown power frame and the power frame of each combination. Finally, the combination achieving the lowest distance is the one that is assigned by the system to the specific frame of the signal. The derivation of the best combination in terms of the distance metric solves the problem of sources enumeration as well as the recognition problem. E.g. if the $\mathbf{s}_1$+$\mathbf{s}_3$ combination is selected the number of sources is 2 and the sources composing the unknown frame are $\mathbf{s}_1$, and $\mathbf{s}_3$.

C. Modeling the classes with GMMs

The second variation employs the Bayesian statistical framework and incorporates the a-priori information we have for the sources in the form of probability density functions of mixture models for each $\boldsymbol{s}_i$. 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 $\mathbf{S}^t = \{\mathbf{s}_1, .., \mathbf{s}_{M(t)}\}$ be the set of sources that compose the observation vector at time *t* according to (4) (e.g. at frame t=10, $\mathbf{S}^{t=10} = \{\mathbf{s}_1, \mathbf{s}_4, \mathbf{s}_5\}$) and $\mathbf{S}_m^t = \{\mathbf{s}_{1,m1}, .., \mathbf{s}_{M(t),m_{M(t)}}\}$ the set of mixtures of the corresponding set of sources that compose the observation vector at time *t* (e.g. at frame t=10, $\mathbf{S}_m^t = \{\mathbf{s}_{1,12}, \mathbf{s}_{4,6}, \mathbf{s}_{5,2}\}$ where the second index is the Gaussian mixture index of the corresponding source). If $\mathbf{S}_m^t$ was known, then from (4):

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

Initially, the observed power vector (where the following cardinality restriction applies: $M(t) \le M, \forall t$) is assumed to be composed of a source $\boldsymbol{s}_i$. This source is selected with uniform probability out of M sources and a mixture $N\left(\mathbf{s}_i; \mu_{i,m}, \Sigma_{i,m}\right)$ (for method A) or a Mel-spectrum (for method B) with probability $w_{i,m}$. 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. In order to predict which and how many sound sources are combined and create the specific mixture, the following approach is adopted: we evaluate a number of H hypotheses where H includes every possible combination of M classes, that is $H=\sum_k M!/(k!(M-k)!)$.

The combination achieving the highest likelihood according to (5) is selected and the sources enumeration and recognition is solved. Please note – and that holds for methods A and B – that enumeration and recognition of multiple sources does not require separation, that is, it not necessary to separate sources by any single-channel technique and then carry out the process of recognition as it is usually the case in literature.

III. Experiments

This section presents the evaluation of the recognition performance of the two methodologies. 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 dog+bird call and the estimated set is dog) and errors in the estimated composition of the set (e.g. sets [bird+rain] and [dog+rain] have the same cardinality but different composition).

During the experimental phase, we employed a great variety of environmental audio signals the sampling rate of which was

16KHz with 16bit analysis. The parameters for extracting the spectral signature out of each signal were the same for method A and B. More specifically, the FFT size was 512, the number of filterbanks was 23, the window size was 25ms and the hop size 12.5ms. With respect to Method B, each source is modeled with a Gaussian mixture of 8 components while the respective parameters (weights, variances and means) were computed using a standard version of the Estimation Maximization algorithm.

The experiments were divided into three categories based on the number of sources: Category A included test signals with one source, Category B signals with two sources and Category C signals with three sources. The audio classes were the following: *bird call, cricket call, dog bark, wind* and *rain noise*. The signals were drawn out of professional sound effects collections (e.g. BBC Sound Effects Library [12]). Signals which contain multiple sources were artificially created by merging equal segments of different sounds at the same energy ratio. The train files which were utilized were the same for both recognition methods so as to have a reliable comparison. Half of the data comprised the train set while the rest served the testing process. Test files were created at a random manner: for example when we wanted to create a mixture with cricket and rain sound events we first selected a cricket sound and subsequently we merged this signal with a portion of a rain sound event of the same size which was chosen randomly. However the files which were used to feed both methodologies were the same.

In Figure 2 we depict the experimental results which were derived using both methodologies. On the top we can see a test sequence which contains test signals from all three categories: initially only a bird call is present. After the first change point a dog barking is added while the last part of the signal is composed of a bird call, a dog barking and a cricket sound event. The recognition results of Methods A and B are demonstrated at the middle and bottom of the figure. The y-axis shows the involved sound sources while a sound source is thought to be active at a given time interval when the particular space on the figure is white. Thorough recognition results will be presented in an enlarged version of the paper as the sources of error are the cardinality as well as the identity of sources themselves; therefore a large number of recognition tests are carried out that due to space limitation are not presented here.

We observe that the segmentation results are very promising as the correct cardinality as well as composition of the mixture is predicted correctly most of the times. The segments which confuse both methods are the onsets and offsets of the sound sources. However, this only shows that a finer portioning of the acoustic signals is needed during the training stage in order to derive more representative spectral signatures (for method A) and models (for method B). There are no algorithms, at least to the knowledge of the authors that can be applied to a relevant task so as to perform comparative experiments. We infer that the statistical modeling technique which is based on GMM is superior to the one which is based on the spectral signature. The GMM approach approximates the patterns of all the involved audio mixtures in a manner which concentrates on the global characteristics of a particular sound class towards limiting the intra-class variability. Thus the trained models have the ability to recognize the novel audio composition accurately and demonstrate robustness in small variations of the spectral content (e.g. different calls of the same bird).

IV. Conclusions

In this work we address the problem of enumeration and recognition of complex sound mixtures using a single microphone. The whole framework is based on the assumption that the audio mix is produced by a subset of sound sources which belong to a known set of classes. Two methods were tested. The first is deterministic and the second probabilistic, that is: a) the first derives the mean power of all possible combinations and b) the second one uses a GMM to fit the power spectrum of every possible combination of the sound sources. Both approaches are based on forming multiple hypotheses on the cardinality and composition of the set of sound sources that is propagated through time. The process acts like a switch and chooses amongst different combinations of sound sources which are a-priori known. We concluded that the GMM modeling technique provides a finer segmentation of the audio sources that exist at a particular soundscene. Our application can be used for the automatic acoustic monitoring of biodiversity and we obtained some quite encouraging recognition results. Future work includes the utilization of these methods on real world data recorded at the Hymettus Mountain for the needs of the AMIBIO project[1]. Furthermore another interesting direction to follow would be the exploitation of temporal information either at feature extraction (e.g. delta coefficients) or during the pattern modeling stage (e.g. hidden Markov models and other Bayesian networks in general).

Acknowledgment

This work was supported by the LIFE+ Program AMIBIO "Automatic Acoustic Monitoring and Inventorying of Biodiversity". We acknowledge the contribution of P. Petrakis spending endless hours in processing bioacoustic signals.

References

[1] O. Wang, D. and G. J. Brown, Computational Auditory Scene Analysis: Principles, Algorithms and Applications, Wiley-Blackwell, Oxford, UK, 2006.

[2] R. Radhakrishnan, and A. Divakaran, "Systematic acquisition of audio classes for elevator surveillance," in Image and Video Communications and Processing 2005, vol. 5685 of Proceedings of SPIE, pp. 64–71, March 2005.

[3] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, and G. Lorho, "Audio-Based Context Recognition", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 321-329, Jan. 2006.

[4] J. Ogle, and D. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," in International Conference on Acoustics, Speech and Signal Processing, Hawaii, pp. I-233-236, 2007.

[1] http://www.amibio-project.eu/

[5] I. Potamitis, "Single channel enumeration and recognition of an unknown and time-varying number of sound sources", in 16th European Signal Processing Conference, Laussane, Switzerland, August 2008.
[6] L. Deng, J. Droppo, and A. Acero, "Estimating Cepstrum of Speech Under the Presence of Noise Using a Joint Prior of Static and Dynamic Features", IEEE Transactions on Speech & Audio Processing, vol. 12, no. 3, pp. 218-233, May 2004.
[7] M. Cowling, and R. Sitte, "Comparison of techniques for environmental sound recognition", Pattern Recognition Letters, vol. 24, no. 15, pp. 2895-2907, Nov. 2003.
[8] F. Sattar, M.Y. Siyal, L.C. Wee, and L.C. Yen, "Blind source separation of audio signals using improved ICA method", 11th IEEE Signal Processing Workshop on Statistical Signal Processing, Singapore, pp. 452-455, 2001.
[9] J. Herre, E. Allamanche, and O. Hellmuth, "Robust matching of audio signals using spectral flatness features," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, pp. 127-130, Oct. 2001.
[10] P. Cano, E. Batlle, E. Gómez, R. De C. T. Gomes, and M. Bonnet, "Audio Fingerprinting: Concepts and Applications", Book Chapter, Springer-Verlag, pp. 233-245, 2005.
[11] E. Allamanche, J. Herre, O. Hellmuth, B. Bernhard Fröbach, and M. Cremer, "AudioID: Towards Content-Based Identification of Audio Material", 100th AES Convention, Amsterdam, May 2001.
[12] http://www.sound-ideas.com/