The Detection of Changes of the Auditory Scene Structure Using a Concept of Short-time ICA

Ladislava Smitkova Janku*, Jiri Smitka**

*Department of Computer Science and Engineering, Czech Technical University in Prague, Czech Republic

**SVM Research
SVM Microwaves, s.r.o, Czech Republic

Abstract

We propose a new method for detecting changes in the scene structure. The detection is based on the evaluation of amount of changes in data acquired from the environment and on the character of these changes. We propose a concept of Short-time Independent Component Analysis and a metrics for comparison of two Generalized Gaussian mixture ICA models.

1. Introduction

A fundamental characteristics of human hearing is the **ability of the selective listening**. Listening to the composite signal originating from the different sound sources, humans are able to separate and identify different sound sources without apparent effort. They are able to separate sound sources despite the fact that many different processes interact to produce the sound signal reaching the ears and that the acoustic events produced by different sources may overlap in time, in frequency or in other characteristics. Cherry [1] used a term of a **Cocktail Party Problem** to denote this human auditory system capability. More recent is a term of **Auditory Scene Analysis (ASA)** investigated by Bregman [2]. The problem can be formulated as follows: we observe a single channel recording which arose as a mixture of different sound sources. The task consists in a separation of these sound sources from the given single channel recording. Source separation can be a parallel process or a sequential one (sequential one is more physiologically plausible). In addition to the conventional techniques (e.g. filtration), the particular attention was given to Computational Auditory Scene Analysis (CASA). While the traditional CASA systems [3] implement less or more grouping rules proposed by Bregman, the systems exploring information theory based approach to CASA use Independent Component Analysis (ICA) [5] or Non-Negative Matrix Factorization (NMF) [4]. From the viewpoint of cognitive psychology, the CASA systems implementing grouping rules are close to the structuralist psychology, while the ICA methods are close to the Gibson’s theory of Directional perception, and particularly to the Information theory of perception [6], [7], [8].

2. State of the Art

ICA has been originally developed as a technique for the separation of the different sources in multi-input multi-output (MIMO) systems. The original sources are estimated from the mixed signals by maximizing their mutual independency. Another application of ICA consists in signal basis decomposition. The basis decomposition of images has been introduced in [8]. The estimation of ICA basis of natural sounds was proposed by Bell and Sejnowski [9]. This approach has been formerly explored in work of [5],[10]. Studying the field of information theory based machine perception, we observe that all ICA-basis decomposition algorithms proposed for single channel source processing (e.g. extracting ICA data-driven basis) process data in a **stationary way**. The set of learned basis filters basis vectors is supposed to be constant for the defined data (image, a set of images, a piece of sound, a set of sound pieces). In [5] the scene analysis is decomposed into IC’s and then the way how to group estimated IC’s together to form sources is discussed. Another important feature, which has not been taking in assumption in [5] is a **high number of ICs forming natural sound scene**. Processing natural sounds using the ICA basis estimation, we usually perform rather projection pursuit than decomposition, because we usually decompose a scene into a smaller number of ICs than is the real number of ICs in the scene. Notice, that increasing number of ICs can improve decomposition results, but cannot assure that the number of computed ICs is equal to the number of ICs in a real scene. Finally, the observed IC’ s of a sound source are dependent on a background. This is a consequence of the limiting factor described above. Real IC’s forming a sound sources can be projected to different IC’s

2.1. Generalized Gaussian Distributions

The generalized gaussian probability distribution (GGD) class are distributions, which are more or less sharp than Normal distribution. Box and Tiao [12] provide a general method for modelling non-Gaussian statistical structure of univariate distributions that have the form \( p(x) \propto \exp(-|x|^q) \); this approach has been explored in Lee and Lewicky [13].

3. ICA Model and Short-time ICA

3.1. ICA basis and ICA basis filters

It is supposed we observe a zero-mean time series \( \tilde{x}(t) = [x(1), x(2), \ldots, x(N)] \), which contains time structure, e.g. the observations in particular time-points are linearly dependent \([x(1), x(2), \ldots, x(N)]\). This time series can be expressed us-
ing data-driven basis as follows:

\[ \tilde{x}(t) = a_1^1(t)s_1 + a_2^2(t)s_2 + \ldots + a_T^T(t)s_T \]  

(1)

\[ \tilde{x}(t) = \sum_{i=1}^{T} a_i^t(s_i) \]  

(2)

where vectors \( \tilde{a}_i, i = 1..k \) represent the vectors of data-driven basis of \( \tilde{x}(t) \) and \( s_1, s_2, s_3 \ldots s_k \) represent the coefficients of \( \tilde{x}(t) \) if it is expressed as a linear superposition of the acquired basis functions \( \tilde{a}_i, i = 1..k \). In order to estimate basis functions sufficiently, more realizations are needed. Processing time-series, we receive more realizations by taking assumption that the probability dependencies among time series components are limited. We assume that the probabilistic dependencies between two samples \( x'(a), x'(b) \) (\( a, b \) denote placements of \( x'(a), x'(b) \) in time) are limited, e.g. there is a natural number \( I \) such that \( x'(a), x'(b) \) are independent if \( |a-b| > I \).

Then, each sample \( x(a) \) belongs to \( I \) different sample vectors except first \( I - 1 \) first samples and last \( N - I + 1 \) samples. We can construct a matrix \( X \) from the samples of an original time series as follows:

\[ X = \mathcal{A}S \]  

(3)

Similarly, we can write

\[ S = BX \]  

(4)

where \( \mathcal{B} \) is inverse or pseudoinverse of \( \mathcal{A} \).

The rank of the basis matrix \( \mathcal{A} = [\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_k] \) (in ICA called also mixing matrix) may be generally \([u_1, u_2], u_1 \neq u_2\), but the ICA algorithms are simplified when we assume \( \mathcal{A} \) to be a square, invertible matrix of a full rank. To simplify computation, we can assume \( I \) is equal to \( K \).

3.2. Generalized Gaussian Mixture ICA Model

ICA Model has been introduced by Lee and Lewicky in [13]. A modelled signal is described as a mixture of ICA basis filters and statistical models of independent components (ICs) or by ICA basis vectors and statistical models of ICs. Because of relation between basis and basis filters, both models are equivalent. Both basis and coefficients (ICs) must be estimated from data. We denote a class of signal realizations \( x(t), t = 1..K, j = 1..C \), by \( \Psi \) and a class model \( M^* \) by

\[ M_{\Psi j} = \{ A_{\Psi j}, p(s_{\Psi j}), l = 1..K \} \]  

(5)

where \( i \) denotes a class, \( j \) denotes a signal realization in a class, and index \( l \) denotes a statistic variable (independent component) \( s^{j*} \). \( p(s_{\Psi j}, l) \) is a pdf of the component \( s^{j*} \) (GGD). We denote by \( N_{\Psi j} \) a number of samples in a class; \( K \) represents a number of independent components and also a number of samples per signal realization.

Similarly, we can define a class model by

\[ M_{\Psi j} = \{ B_{\Psi j}, s_{\Psi j}, l = 1..K \} \]  

(6)

where \( B_{\Psi j} \) denotes estimated 'basis filters' (an estimated basis filter matrix or an estimated demixing matrix in ICA terminology). In opposite to Lee and Lewicky, who use Natural gradient algorithm with switching nonlinearity, we use sequential extraction of independent components [14]. This approach may have the following advantages over simultaneous blind separation: signals can be extracted in a specified order according to the stochastic features of the source signal, and the learning algorithms are local and biologically plausible. The learning algorithm is based on FASTICA [15]. The data is centered and whitened before ICA algorithm application. We describe the method in [16] in detail.

3.3. A Concept of Short-Time ICA

In this section, we introduce a concept of short-time ICA (STICA). The idea is similar to the short-time processing in a time or Fourier domains. The input signal is divided into overlapping segments. These segments are processed by the ICA algorithm separately. We estimate the data-driven basis filters for each segment.

To compare two ICA models, we limited the task of model comparison to the signal segments, which are overlapped in the way as it is shown on Fig. 1. Each model \( M' \) is perceived as some updated version of any ICA model \( M \) estimated before \( j < i \). We introduce the metrics, which is based on Kullback-Leibler divergence.

Diversity between two ICA models can be defined as:

\[ D_m(W_i, W_j) = K(W_i, W_j) + L(B_i, B_j) \]  

(7)

If we omit the member related to the diversity of basis filters, we can express this measure as the sum of Kullback-Leibler divergencies between couples of GGD describing independent components:

\[ D_m(W_i, W_j) = K(W_i, W_j) = \sum D(p(x; \alpha_1 \beta_1)||p(x; \alpha_2 \beta_2)) \]  

(8)

However, such omitting basis filter structure is possible only if the overlapping segment structure shown on Fig. 1 is used for signal analysis.

For the following GGD family

\[ p_{\beta}(x; \alpha), \alpha \in \mathbb{R}^+ : p_{\beta}(x; \alpha) = \frac{\beta}{2\alpha \Gamma(1/\beta)} e^{-(x/\alpha)^{\beta}} \]  

we have the

\[ D(p(x; \alpha_1 \beta_1)||p(x; \alpha_2 \beta_2)) = \log \left( \frac{\beta_1 \alpha_2 \Gamma(1/\beta_1)}{\beta_2 \alpha_1 \Gamma(1/\beta_1)} \right) + \left( \frac{\alpha_1}{\alpha_2} \right)^{\beta_2} \frac{\Gamma((\beta_2 + 1)/\beta_1)}{\Gamma(1/\beta_1)} - \frac{1}{\beta_1} \]  

(9)
the comparisons of the following model couples are computed:

\[
M_{ij} = \text{pared pairwise to the parameters of } j_i \text{ segments does stay constant. In each step length of the segment window nor the length of the overlapping}
\]

Fig. 2. In opposite to the standard short-time methods, nor the characters.

What is important is an amount of these changes and their characteristic when the auditory scene consists only of one sound source.

In this section, we introduce a new algorithm for the detection of changes of the structure of an auditory scene (source adding or taking away). There are three fundamental ideas behind this method:

- The auditory scene analysis is a dynamical process. Sounding objects are detected on a basis of changes in the sound environment. The development of a scene in time is an important factor in scene analysis.
- In mixture, each sounding object is perceived against background. This background is formed by other sources forming the mixture.
- The detection of each scene structure change is based on the amount of changes in data acquired from the environment and on the character of these changes.

Naturally, the changes in environmental data are detectable also when the auditory scene consists only of one sound source. What is important is an amount of these changes and their characters.

The demonstration of the proposed method is shown on Fig. 2. In opposite to the standard short-time methods, nor the length of the segment window nor the length of the overlapping segments does stay constant. In each step \( i \), the parameters of the ICA model \( M_i \) are estimated. These parameters are compared pairwise to the parameters of \( j \) ICA models constructed previously (\( j \) is a natural, finite number). In other words, the the comparisons of the following model couples are computed:

\[
M_{i}A \rightarrow M_{i-1}A, \quad M_{i}A \rightarrow M_{i-2}A, \ldots, \quad M_{i}A \rightarrow M_{i}A.
\]

These comparisons result in a set of values \( \Delta M_{ij} \) which characterize mutual model differences. In each step, we analyze the acquired set of mutual model differences (MMDs). The value of MDD criterion is computed on a basis of Eq. 8 (as a weighted sum). The structure of the MMD set describes the changes in auditory scene between time points \( i \) and \( i - j \).

The experimentally acquired MMD for a single channel recording consisting of one sound source (speech utterance, female) is shown on Fig. 3 (a). For the demonstration purposes, the pauses between words have been extracted from the speech utterance. Naturally, we observe the smallest MMD \( \Delta M_i^{i-1} \) (MMD between the model \( i \) and \( i - 1 \)) and the largest MMD \( \Delta M_{i} \) (MMD between the model \( i \) and \( j \)). The results are similar for male voices. The results changed rapidly, when another sound source appeared in the scene in a time interval \( (i - j, i) \). See Fig. 3 (b) for the demonstration. We use two speech utterances from the TIMIT database (male voice, female voice). For the demonstration purposes, the pauses between words have been extracted from the speech utterances.

The sound source B appeared at time point \( k \). The change in MDD curve is not satisfactory detectable immediately after the sound source appeared. Some ’integration’ time is needed.

To be able to separate and finally to track the new sound source, we incorporate an algorithm for maximum-likelihood estimation of a single sound source from mixture proposed by Jang and Lee [10] in our method. At the time-point of change detection \( k + a \), the model of the sound source \( A \) is available. Using MDD curves, we can detect the optimal time point, at which the ICA model of source \( A \) is not corrupted by the source \( B \) (or it is only slightly corrupted). This ICA model of source \( A \) is used to decompose observed mixture into source \( A \) and newly detected sound source \( B \) using method described in [10]. ICA model parameters of the part of the signal separated as a source \( B \) are estimated. Simultaneously, the the system tracks mixture \( A+B \) and estimates the ICA model of \( A+B \). The proposed method is recursive - if the third sound source \( C \) is added to the mixture, it is separated against background \( A+B \).

The change in MMD reflects the change in auditory scene (both the start and the end of a new sound source). To be able to detect the character of the change, we use: 1) Overall energy. 2) A comparison of ICA model of the recording to ICA models estimated previously.

If the third sound source \( C \) is added to the mixture \( A+B \), the change in a scene structure is detected. The system computes the overall energy of the signal. It computes the ICA model of the recording \( (A+B+C) \) and compares it to the known ICA models (ICA model of the source \( A \), ICA model of the source \( B \)). If the sound source \( B \) consisted of two independent sources \( B_1 \) and \( B_2 \) starting simultaneously, the model would not be able to separate them without knowledge on \( B_1 \) and \( B_2 \) learned in the past. The proposed method would have to be extended by incorporating comparison of newly detected sound source ICA model with combinations of saved ICA models (notice, that it is out of the scope of this paper).

To improve the stability of the proposed method, we use a concept of basic background. The basic background is a random noise of a very low intensity added to the mixture. A simple sound source is processed against this basic background.

5. Experiments

We test our algorithm on auditory scene investigated in experiment proposed by Bregman and Pinker [14]. This experiment shows the activity of the scene analysis process - the streaming effect and the illusory continuation of one sound behind another. We want to demonstrate that our grouping algorithm achieves similar results as humans do.

The sounds used in this experiment are shown in Fig. 4.
posed method: an cumulative analysis of amount and structure of changes in data. The method consists in the idea of the source detection by the accu-
novation to the novel concept of STICA, the novelty of the presented
analysis the amount and structure of changes in data. In addi-
CASA, our method does not explore grouping cues; rather it
in the environmental data. In comparison to the traditional
source detection based on the analysis of accumulated changes
in ICA (STICA) basis decomposition and the new method of

In this paper, we propose both the new concept of Short-time
ICA (STICA) basis decomposition and the new method of
source detection based on the analysis of accumulated changes
in the environmental data. In comparison to the traditional
CASA, our method does not explore grouping cues; rather it
in the environmental data. In comparison to the traditional
source detection based on the analysis of accumulated changes
in data. In additional to the novel concept of STICA, the novelty of the presented
method consists in the idea of the source detection by the accu-
mulative analysis of amount and structure of changes in data.

We summarize the most important properties of the pro-
posed method:

- Sounding objects are detected on a basis of changes in
the sound environment.
- Each sounding object is perceived against background. This
background is formed by other sources forming the
mixture.
- The basic background is a random noise of a very low
intensity. A simple sound source is processed against
this basic background. This basic background concept
has been added to the method to increase its stability.
- Detecting sound, this method can explore only information
acquired from data. Both information present in a
processed data (current sound environment) and information
acquired from data in the past, saved in some
form of knowledge (e.g. models).

7. Acknowledgements
The research of Jiri Smitka has been supported by the project
FI-IM 145/2 of the Ministry of Industry and Trade of Czech.
Rep.

8. References

25(53) 975-979
[4] Lee, D.D., and H.S. Seung, Learning the parts of objects by
achusetts.
maximization approach to blind separation and blind
[7] Olshausen, B.A., and Field, D.J., Emergence of Simple-
Cell Receptive Field Properties by Learning a Sparse Code
Component Analysis. John Willey and Sons. "01.
Structure of a Nat-ural Sound. Network: Computation in
Neural Systems, '96(7).
proach to Single-channel Source Separation", Journal of
and Image Processing, John Willey and Sons, New York.
[12] Box,G. and Tiao,G., Bayesina Inference in Statistical
Analysis, John Wiley and Sons.
sian Mixture Model Using ICA," IEEE Transactions on Image
for Independent Component Analysis", IEEE Trans. Neural
Network, 10(3).
[16] Smitkova Janku, L., Learning Algorithms Based on High-
order Statistic: A Unifying Framework for Analysis of
Mixtures of Natural Sounds, Ph.D.Thesis, Czech Technical
University in Prague.