Computational acoustics and psychoacoustics in spatio-temporal simulated segregation of sound sources

Theses of the Ph.D. Dissertation

Zoltán Fodróczi

Péter Pázmány Catholic University
Faculty of Information Technology
Multidisciplinary Technical Sciences Doctoral School

Hungarian Academy of Sciences
Computer and Automation Research Institute
Analogic and Neural Computing Laboratory

Scientific Advisor :
Radványi András D.Sc
Doctor of the Hungarian Academy of Sciences

Budapest, 2007

1 Introduction

At the beginning of the 21st century we use basically the same keyboard that first was used to input data to Binac[1] in 1948. Since 1964 when the mouse was invented by Doublas Engelbar[2] there has been no remarkable change in the human-computer interaction. However one of the elementary visions of most of the sci-fi writers is that our computers and other electronic gadgets can be controlled through our voice commands.

Thanks to the result of the last few decades of research, we are already in possession of algorithms that can recognize our voice with comparable performance to the human listeners. Unfortunately this fact holds only for noise free conditions. In noisy cases the performance of artificial systems are dramatically decreasing, since the algorithms can not handle signals not belonging to the recognition task.

Towards the solution of this problem, technological advances in sound analysis are needed in order to solve several basic problems, such as speaker localization and tracking, speech activity detection or automatic speech recognition of distant talk. The long-term goal is the ability to monitor speakers and noise sources in a real reverberant environment, with acceptable constraint on the distribution of microphones or on the number of active sound sources. This problem is surpassingly difficult [1], as the speech signals collected by a given set of microphones are severely degraded by both background noise and reverberation. The performance of the algorithms above crucially depends on the pre-processing stage where relevant information is extracted from the noisy and complex mixture of audio signals produced by several members of the acoustic scene.

Living organisms exhibit a remarkable ability to solve sound segregation problem through separating sound objects of interest from a mixture of other intruding sounds. This phenomenon may be regarded as an aspect of a more general process of auditory system, which is able to untangle an acoustic mixture in order to retrieve a perceptual description of each constituent sound source. The term auditory scene analysis (ASA) has been introduced to describe this process. Conceptually, ASA may be regarded as a set of several stages. I consider the problem of separation of sound signals in everyday acoustic environment. I divide the problem into two parts. First, I discuss segregation based on physical properties of sound signals according to the heuristic algorithms known from psychoacoustics. I introduce a wave computer implementation of the grouping rules of Auditory Scene Analysis (ASA) [2]. Based on the properties of the applied Cellular Neural Network (CNN) Universal Machine architecture [3] this implementation makes fast and efficient computation of several grouping rules possible and opens the door of real time applications.

I also discuss the problem of sound source localization which is considered as a method of sound source segregation based in spatial cues. I introduce a novel algorithm that makes possible the localization of anisotropic sources in reverberant environment by takig into account the effect of the acoustic environment.

2 Methods of Investigation

In the course of my work, theorems and assertions from the field of mathematical statistics, psychoacoustics, computational acoustics, cellular neural networks, machine learning and signal processing were utilized. During my studies, I developed an efficient cochlea simulator that preserves and

[1] Binac was the first computer that has an other input device apart from the punched card reader. This device was the keyboard of a typewriter. [http://inventors.about.com]

[2] Inventor of the mouse we use. [http://inventors.about.com]

emphasizes the information content of sound signals required by the proposed implementation of auditory scene analysis. A two-dimensional spectro-temporal flow was created by the simulator, which was utilized by the auditory wave computing library. I gave the high level, "CNN type language" description (Universal Machine on Flows), and the low level, hardware close AMC language description of the algorithms included in the library. During designing the algorithms I paid attention to compatibility issues between different hardware and software platforms. An important aspect was the use of the locally connected linear and robust templates in order to be able to run the algorithms on the current CNN-UM VLSI chips. Stability and robustness of the new template class I developed were examined by the tools of mathematical analysis and by the theorems of CNN theory. The analogic algorithms were tested on the AladdinPro software simulator. Utilities required by cross-platform applications were created using the Matlab development environment. For the problem of spatial identification of sound sources I defined a model based on the specular reflection method. I examined the limitations of the model, and by utilizing the tools of mathematical analysis I concluded the consequences of anisotropic sources in reverberant environments. Using the mathematical probability I made possible to predict the effect of acoustic environment to the cross-correlation function. Based on the experiences of machine learning I gave a method to measure the similarity between predicted reverberation effect and the observations gathered by measurements. The proposed method was implemented in C++. This implementation was used to test the performance of the algorithm in a simulated acoustic environment that was created applying the CATT[3] acoustic modeling software.

3 New scientific results

Thesis group 1.
I devised the Auditory Wave Computing Toolkit (AWCT) a Computational Auditory Scene Analysis library. The algorithms provide an efficient tool to mimic some aspects of the human auditory system. The toolkit contains an efficient sound-to-image transformation method that is analogue to the cochlea in its functioning and provides two-dimensional spectro-temporal flow and contains the implementation of a set of psychoacoustic grouping procedures known from the field of Auditory Scene Analysis.

In the first stage of the auditory system, the acoustic mixture is decomposed into sensory elements. The sensory elements are grouped together to form "groups" of different hierarchy. At the end of the process low level or primitive groups combine auditory objects, which are likely to have originated from the same sound source. Recently, attempts to develop computational systems that mimic ASA have led to the emergence of a new field, known as computational auditory scene analysis (CASA). A key feature of these algorithms is that they process spatio-temporal trajectories i.e. flows. The Cellular Wave Computer on Flows realized on the CNN Universal Machine (CNN-UM) is a promising candidate for efficiently implementing CASA rules, due to its layered cellular prototype architecture. This thesis presents the CASA CNN-UM program library called Auditory Wave Computing Toolkit library (AWCT library), which implements CASA rules on CNN-UM.

1.1. The base for the common onset rule is that if a particular physical process generates energy in several frequency bands, it is likely that energy flow will start in each band at the same moment. This regularity in our auditory environment has introduced an interpretation

[3] Computer Aided Theater Technique

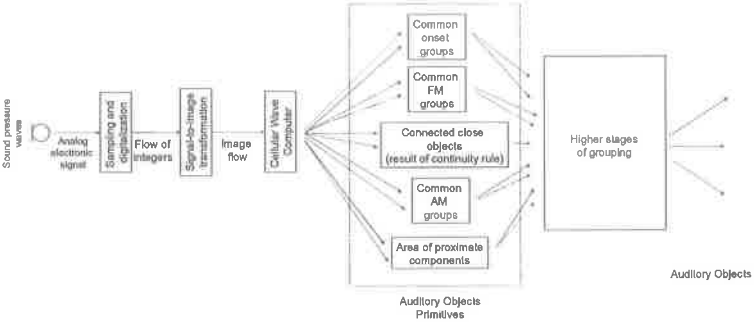
Sound pressure waves
Analog electronic signal
Sampling and digitalization
Flow of integers
Signal-to-image transformation
Image flow
Cellular Wave Computer
Common onset groups
Common FM groups
Connected close objects (result of continuity rule)
Common AM groups
Area of proximate components
Auditory Objects Primitives
Higher stages of grouping
Auditory Objects

Figure 1: The Auditory Wave Computing Toolkit.

of "common onsets" between components, as an indication that these components originate from the same source. I developed a new wave computing algorithm for the implementation of "common onset" rule. The algorithm detects synchronous energy bursts in the two-dimensional spectro-temporal map based on logical operations and collision of propagating cellular waves.

1.2. The common fate i.e. common amplitude/frequency modulations are the major and most important cues for auditory grouping. Component groups exhibiting these commonalities are, generally, perceived as one object, making it hard for the human listener to focus on any individual components. Human behaviour under these cues is attributed to the fact that physical sound producing systems, under some manipulation modulating their frequency or amplitude characteristics, will imprint this modulation in all their auditory components.
Common amplitude modulation originates in synchronous onset and offset. Based on the results of thesis 1. the previous point I gave a method to detect these commonalities by exploiting the computational capabilities of cellular waves.
The common frequency modulation in a log-frequency scale is displayed as energy traces parallel to each other. The proposed algorithm exploits this feature and detects the common spectral distance using a new NxN robust CNN template class. Members of this template class can be decomposed in linear steps into a sequence of 3x3 template operations making possible the application on silicon implementations of CNN-UM.

1.3. Acoustic energy in a certain frequency band originating from a single source may undergo brief interruptions or periods of undetectability after which it will reappear close to the frequency in question. When the "gaps" are short in duration compared to the energy bursts, it is desirable to make some kind of connection across the gaps to indicate the similarity and closeness between two parts. I introduced a method that is able to fill the "gaps" caused by this factor resulting in continuous sound objects.

1.4. Accroding to the proximity grouping principle the closer the energy bursts are to each other in time and frequency the more likely it is that they can be fused together. I developed a new method that exploits the capabilities of the applied cellular architecture. It marks regions of higher average energy than a given threshold resulting in the objects which can be grouped by the proximity principle.

Published in:

***Z. Fodróczi**, A. Radványi "Computational Auditory Scene Analysis in Cellular Wave Computing Framework" International Journal of Circuit Theory and Applications Vol: 34(4) pp: 489-515, ISSN:0098-9886 (July 2006)*

Thesis group 2.
I developed a new sound source localization algorithm which makes possible localization of anisotropic sound sources in reverberant environment. The method applies the specular reflection model of sound reflection to integrate the effect of the directivity of sound sources. This effect is considered during localization supporting real-time reverberation tolerant source localization without the need of special hardware system.

Speaker localization with microphone arrays has received significant attention in the past decade as a means for automated speaker tracking of individuals in a closed space for videoconferencing systems, directed speech capture systems and surveillance systems. Traditional techniques are based on estimating the relative time difference of arrivals (TDOA) between channels, by utilizing cross-correlation function.

Earlier studies on source localization have not considered the directional characteristics of the source, however, by examining the effect of source directivity, several phenomena can be explained. The relatively weak performance of the currently used TDOA based speaker localization systems is interpreted as the consequence of reverberation that causes spurious peaks in the cross-correlation function, since two reflected paths with the same propagation delay to the microphone may add leading to a higher peak, resulting in false TDOA estimation. By taking source and microphone directivity into account, the coincidence of time difference of reverberation paths is not a necessary condition for the occurrence of false TDOA estimation. Due to the joint effect of the source and microphone directivity, a less attenuated reverberation path may result in a peak higher than that of the direct path.

The method I propose utilizes a priori acoustic information of the monitored region and makes possible to localize directional sound sources by taking the effect of reverberation into account.

2.1. According to the applied sound reflection model, I developed a mathematical formula to describe the time function of signals recorded by microphone in reverberant environment. Using this formula I showed that the cross-correlation function can be given as the linear combination of shifted auto-correlation functions. By examining the properties of the auto-correlation function I gave a method to predict the effect of reverberation on the cross-correlation function.

2.2. I analyzed the effect of anisotropic sound source on the cross-correlation function. I identified the circumstances when the joint effect of source directivity and reverberation causes the traditional TDOA methods fail.

2.3. I adapted the *accumulated correlation method* to create the predicted-reverberation-effect-maps serving an efficient and robust method to integrate the predictions of different microphone pairs. I defined a four-dimensional point set that is a characteristic identifier of a predicted-reverberation-effect-map and a given acoustic configuration.

2.4. I introduced a method to extract the effect of reverberation from the *accumulated correlation map*, and I gave a method to find the best matching pre-stored configuration to observation. The best matching configuration is the hypothetical source location.

Published in:

*Z. **Fodróczi**, A Radványi. "Localization of Directional Sound Sources Supported by a priori Information of the Acoustic Environment" manuscript submitted to EURASIP Journal on Applied Signal Processing*

4 Application of the results

The results introduced in the first thesis can be utilized as a pre-processing stage that aims is to provide independent sound objects originating from a single source. Such an application could be signal pre-selection of a sound source localizer. The segregated objects can be further used in automatic speech/sound recognition systems. The properly selected computational architecture makes possible the application of portable systems such as a hearing aid or a cochlear implant where the need of high computing power with low energy dissipation is essential to build high quality adaptive context sensitive gain control.

The results introduced in the second thesis make possible a more reliable speaker localization in reverberant environment. This advantage can directly increase the performance of security surveillance systems or the seamless operation of tracking capabilities of videoconferencing systems. With an additional directed microphone array, the knowledge of speaker's location could help to make better records than previously, which could indirectly decrease the false word recognition rate of automatic speech recognition systems.

5 The way of further research

Methods introduced in the first thesis group are similar to the data driven or primitive grouping procedures of the human hearing system. However, in our perception the schema driven procedures bear also a crucial role. The information is conveyed from higher stages based on our previously learnt experience. Such an experience could be the location information and the Doppler-effect, for instance, in the segregation of the sound of a passing car. The previously learnt pattern and the predictable behaviour of known sound signals have a more significant role in our perception. Thank to this schema driven procedures the recognition problem can be solved on a limited domain of possible events. This domain is developed according to the context of a previously known sound scene or by the support of our other perceptions. This function is responsible for our capability to recognize speech in high background noise. Now days it is not clear how the schema driven mechanisms alter the evaluation of data driven rules. It is likely that evaluation of this rules are also controlled by some predictive model. It is a remarkable hypothesis in cognitive neuroscience that the *event related potential (ERP)* which can be measured by a set of EEG electrodes is a

marker of the update of predictive models. We already have a quite detailed knowledge about the characteristic features of ERP so it is opportune to develop a functionally analogue artificial system which can form an important part of future sound processing systems.
From the results proposed in the second thesis group follow that source localization can not be solved simply by the determination of time delay of arrival, since the anisotropic properties of the sound source and the joint effect of reverberation yields to bad estimations. It is indispensable to integrate or to filter out the effect of acoustic environment. In my study I showed an example to the integration. The introduced method is not perfect, it is recommended to develop the integration of the effect of reverberation through global cues in order to increase the noise sensitivity. The application area could be extended by the usage of frequency dependent predictions. It would make choosing the most fitting prediction set possible, based on the spectral content of the incoming signal.
It is likely that in case of biological system the learnt effect of the body to the acoustic environment is an important cue. When an animal turns its head, the alteration of the spectral content of the incoming sound can be used to find out the direction of the sound source. Until now there is no any artificial system that would take this pro-active strategy into account.

6 Acknowledgements

The first acknowledgement shall be entitled to Professor Tamás Roska, the leader of the Analogic and Neural Computing Laboratory at the Computer and Automation Research Institute of the Hungarian Academy of Sciences and the dean at the Information Technology Department of Pazmany Peter Catholic University. He supported the crucial financial and professional background to my studies.

I am also thankful to Dr. György Takács, who was my supervisor at the beginning of my studies and introduced me to the world of acoustics. He supported me with his advices during the whole time of my work that helped me a lot to find the way in the realm of signal and speech processing.

Dr. Péter Szolgay was my undergraduate teacher at Veszprém University. He showed me the gate to the scientific work and supported me with his friendship.

Dr. Bércesné Novák Ágnes was the lecturer of the practical courses I gave at Pazmany Peter Catholic University. I thank her in addition my academic duty I had enough time to do my research. Her endless advocacy was really important.

I also thank to Dr. András Illényi, the professor of the Budapest University of Technology and Economics. He opened the door of the Békésy György Acoustic Laboratory where I was able to make some irreplaceable experiments.

I am also thanking for all of my laboratory co-workers for helping in research and study in various kind and for the good atmosphere in the Analogic laboratory and at the Pazmany University. Especially I thank among them Attila Kis, Róbert Wágner, Péter Jónás, Csaba Benedek, Barnabás Hegyi, László Havas, Gábor Vásárhelyi, Anna Lázár, Anrás Bárdi, Tamás Harczos and Gergő Feldhoffer.

I spent six month in Germany at the Fraunhofer Institute in Ilmenau. I met there several nice people such as Dr. Frank Klefenz, András Kátai, Stephan Werner and Wolfgang Köstritzer. They helped me a lot not just professionally but also with their humanity.

I am grateful to Éva Nóra Nagy, János Korn and Márton Péri for helping me to increase the quality of English in my publications.

The economical and academic support of the Computer and Automation Research Institute of the Hungarian Academy of Sciences was really important. I found there always a nice and inspirative atmosphere that helped my work.

The role of my advisor in the success of my work is not expressible. He supported me at every stages of my work. But for him I would have not been able to accomplish my studies.

This way of acknowledgement is not the right to tell what I got from my family and friends.

7 The author's publications

Z. Fodróczi, A. Radványi "Computational Auditory Scene Analysis in Cellular Wave Computing Framework" International Journal of Circuit Theory and Applications Vol: 34(4) pp: 489-515, ISSN:0098-9886 (July 2006)

Z. Fodróczi, A Radványi. "Localization of Directional Sound Sources Supported by a priori Information of the Acoustic Environment" manuscript accepted to EURASIP Journal on Applied Signal Processing

Z. Fodróczi, A. Radványi, Gy. Takács "Acoustic Source Localization using Microphone Arrays via CNN algorithms" Proceedings of 3rd International Conference on European Conference on Circuit Theory and Design (ECCTD03) 2003

References

[1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. *Microphone Arrays: Signal Processing Techniques and Applications.* Springer, New York, NY, USA, 2001.

[2] Albert S. Bregman. *Auditory Scene Analysis.* MIT Press, Cambridge, 1990.

[3] T. Roska and L. O. Chua. The CNN Universal Machine: an Analogic Array Computer. *IEEE Transactions on Circuits and Systems-II*, 40:163–173, 1993.