IEEE

The Institute of Electrical and Electronics Engineers Inc.
Region 8 – Europe, Middle East and Africa
Poland Section, Chapter Signal Processing, Circuits and Systems

Audio Engineering Society
Polish Section

PROCEEDINGS OF THE JOINT CONFERENCE:

NTAV/SPA 2012

NEW TRENDS IN AUDIO AND VIDEO
SIGNAL PROCESSING: ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

ŁÓDŹ, POLAND, 27-29 SEPTEMBER 2012

Lodz University of Technology
Institute of Electronics

wydział
elektrotechniki
elektroniki
informatyki
i automatyki

Polish Society
of Theoretical and Applied
Electric Engineering

NTAV/SPA 2012

NEW TRENDS IN AUDIO AND VIDEO SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

CONFERENCE PROCEEDINGS

ŁÓDŹ, 27-29 SEPTEMBER 2012

Conference Organizer

Lodz University of Technology
Faculty of Electrical, Electronic, Computer and Control Engineering
Institute of Electronics, Medical Electronics Division
211/215 Wólczańska Str., B-9 building, 90-924 Łódź, Poland

Co-organizers

Poznan University of Technology
Faculty of Computing Science and Management
Chair of Control and System Engineering
Division of Signal Processing and Electronic Systems
3 Piotrowo Str., 60-965 Poznań, Poland

Wroclaw University of Technology
Institute of Telecommunication, Teleinformatics and Acoustics
Department of Acoustics
27 Wybrzeże Wyspiańskiego Str., 50-370 Wrocław, Poland

Wrocław
University
of Technology

Polish Society of Theoretical and Applied Electric Engineering
18/22 Stefanowskiego Str., 90-924 Łódź, Poland

Faculty of Electrical, Electronic, Computer and Control Engineering

The conference was financially supported by the Polish Ministry of Science and Higher Education

**Ministry of Science
and Higher Education**
Republic of Poland

Published by:
Lodz University of Technology Press

Edited by:
Paweł Strumiłło, Michał Bujacz, Monika Popielata

Cover design:
Anna Janisz

ISBN 978-83-7283-502-4
NTAV-SPA 2012 Conf ID 21037

Printed by: QuickDruk, 11 Łąkowa Str., Łódź

Audio Engineering Society IEEE PTETiS

NTAV/SPA 2012

NEW TRENDS IN AUDIO AND VIDEO SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

CONFERENCE PROCEEDINGS

ŁÓDŹ, 27–29 SEPTEMBER 2012

Scientific Committee

CHAIRMAN	
Prof. Andrzej Materka	Lodz University of Technology, Poland
VICE-CHAIRMAN	
Prof. Adam Dąbrowski	Poznan University of Technology, Poland
VICE-CHAIRMAN	
Prof. Andrzej Dobrucki	Wroclaw University of Technology, Poland
Prof. Andrzej Bartoszewicz	Lodz University of Technology, Poland
Dr. Michał Bujacz	Lodz University of Technology, Poland
Prof. Ryszard Choraś	University of Technology and Life Sciences Bydgoszcz, Poland
Prof. Zygmunt Ciota	Lodz University of Technology, Poland
Prof. Anthony Davies	King's College London, Great Britain
Prof. Marek Domański	Poznan University of Technology, Poland
Prof. Andrzej Dziech	AGH University of Science and Technology Cracow, Poland
Prof. Alfred Fettweis	Ruhr-Universität Bochum, Germany
Prof. Ewa Hermanowicz	Gdansk University of Technology, Poland
Prof. Lidia Jackowska-Strumiłło	Lodz University of Technology, Poland
Prof. Michał Jacymirski	Lodz University of Technology, Poland
Prof. Jacek Kabziński	Lodz University of Technology, Poland
Prof. Tomasz Kacprzak	Lodz University of Technology, Poland
Prof. Mos Kaveh	President of the IEEE Signal Processing Society
Prof. Piotr Kleczkowski	AGH University of Science and Technology Cracow, Poland
Prof. Christian Kollmitzer	University of Applied Sciences Technikum Wien, Austria
Prof. Bożena Kostek	Gdansk University of Technology, Poland
Prof. Krzysztof Kozłowski	Poznan University of Technology, Poland
Prof. Rolf Kraemer	IHP Microelectronics, Frankfurt/Oder, Germany
Prof. Jacek Kucharski	Lodz University of Technology, Poland
Prof. Zbigniew Kulka	Warsaw University of Technology, Poland
Prof. Józef Modelski	Warsaw University of Technology, Poland
Prof. George Moschytz	Bar-Ilan University, Israel
Prof. Włodzimierz Mosorow	Lodz University of Technology, Poland
Prof. Andrzej Napieralski	Lodz University of Technology, Poland
Prof. Peter Noll	Technische Universität Berlin, Germany
Prof. Antoni Nowakowski	Gdansk University of Technology, Poland
Prof. Maciej Ogorzałek	Jagiellonian University in Krakow, Poland
Prof. Stanisław Osowski	Warsaw University of Technology, Poland
Prof. Aleksander Petrovsky	Bialystok University of Technology, Poland
Prof. Jan Purczyński	West Pomeranian University of Technology, Szczecin, Poland
Prof. Kamisetty R. Rao	University of Texas at Arlington, USA

Dr. Agnieszka Roginska	New York University, USA
Prof. Marek Rudnicki	Lodz University of Technology, Poland
Prof. Milan Rusko	Slovak Academy of Sciences, Bratislava, Slovak republic
Prof. Aleksander Sęk	Adam Mickiewicz University Poznan, Poland
Prof. Thomas Sikora	Technische Universität Berlin, Germany
Prof. Wladyslaw Skarbek	Warsaw University of Technology, Poland
Prof. Paweł Strumiłło	Lodz University of Technology, Poland
Prof. Michał Strzelecki	Lodz University of Technology, Poland
Prof. Piotr Szczepaniak	Lodz University of Technology, Poland
Prof. Krzysztof Ślot	Lodz University of Technology, Poland
Prof. Ryszard Tadeusiewicz	AGH University of Science and Technology Cracow, Poland
Prof. Ralph Urbansky	Universität Kaiserslautern, Germany
Prof. Joos Vandewalle	Katholieke Universiteit Leuven, Belgium
Prof. Heinrich T. Vierhaus	Brandenburgische Technische Universität Cottbus, Germany
Dr. Gyorgy Wersenyi	Széchenyi István University, Hungary
Prof. Ryszard Wojtyna	University of Technology and Life Sciences Bydgoszcz, Poland
Prof. Jan Zarzycki	Wroclaw University of Technology, Poland

Conference Organizing Committee

CHAIRMAN

Prof. Paweł Strumiłło	Lodz University of Technology, Poland

VICE-CHAIRMAN

Michał Bujacz, Ph.D.	Lodz University of Technology, Poland

Members:

Julian Balcerek, M.Sc.	Poznan University of Technology, Poland
Anna Borowska-Terka, M.Sc.	Lodz University of Technology, Poland
Agnieszka Chmielewska, M.Sc.	Lodz University of Technology, Poland
Wanda Gryglewicz-Kacerka, Ph.D.	Lodz University of Technology, Poland
Barbara Kociołek	Lodz University of Technology, Poland
Aleksandra Królak, Ph.D.	Lodz University of Technology, Poland
Tomasz Marciniak, Ph.D.	Poznan University of Technology, Poland
Bartosz Ostrowski, M.Sc.	Lodz University of Technology, Poland
Paweł Pawłowski, Ph.D.	Poznan University of Technology, Poland
Przemysław Plaskota, Ph.D.	Wroclaw University of Technology, Poland
Aleksandra Sibińska, M.Sc.	Lodz University of Technology, Poland

IEEE

CONFERENCE INFORMATION

Conference Record #: 21037

Conference Title: 2012 Joint Conference New Trends In Audio & Video And Signal Processing: Algorithms, Architectures, Arrangements And Applications

Conference Acronym: NTAV/SPA 2012

Conference Dates: 9/27/2012 to 9/29/2012

Location: Lodz University of Technology

City: Lodz

Country: Poland

Exhibits: N **No. of Exhibits:** 0 **Tutorials:** Y **Attendance:** 100 **Producing Publication:** Y **Concentration Banking Info:** No

Conference Scope: DSP theory, algorithms and implementations, Image and video processing, Audio processing, Digital Television and stereovision, Multimedia data compression and editing, data bases, Electroacoustics, Psychoacoustics, Sound mastering, Human computer interfaces, Image synthesis, analysis and recognition, Filter design and implementation, Vision and audio-based diagnosis in medicine and industry, Text-to-speech & speech-to-text algorithms, Artificial intelligence applications in DSP, Distributed and networked DSP systems, Virtual and augmented reality, Biometrics

Conference Keywords: dsp, image processing, audio processing, speech processing, multimedia, audio and visual systems, ntav/spa 2012

Conference Focus: Application, Scientific/Academic

WWW URL: http://www.ntavspa.pl

CFP URL: http://www.ntavspa.pl/__files/call_ntav_spa_2012.pdf

Abstract Submission Date: 5/31/2012

Notification of Acceptance: 7/15/2012

Final Paper Submission Date: 8/15/2012

Expenses: -

Revenue: -

Sponsors

Lodz University of Technology, Poland (Co Sponsor) – 40%

Audio Engineering Society – AES (Co Sponsor) –30%

IEEE Poland Section Signal Processing Chapter (Co Sponsor) –10%

IEEE Poland Section Circuits and Systems Chapter (Co Sponsor) –10%

Poznan University of Technology, Poland (Co Sponsor) –10%

Information Contact:
Julian Balcerek
Piotrowo 3
Poznan University of Technology
Chair of Control and System Engineering
Division of Signal Processing and Electronic Systems
Poznan 60-965
POLAND
Ph : +48 61 665 2833
Fax: +48 61 665 2840
julian.balcerek@put.poznan.pl

Technical Program Chair:
Paweł Strumiłło
211/215 Wolczanska Str., B-9 building
Lodz University of Technology
Lodz 90-924
POLAND
Ph : +48 42 631 2646
Fax: +48 42 636 2238
pawel.strumillo@p.lodz.pl

Conference Chair:
Andrzej Materka
211/215 Wolczanska Str., B-9 building
Technical University of Lodz
Lodz 90-924
POLAND
Ph : +48 42 631 2644
Fax: +48 42 636 2238
andrzej.materka@p.lodz.pl

Publication Chair:
Tomasz Marciniak
Piotrowo 3
Poznan University of Technology
Chair of Control and System Engineering
Division of Signal Processing
and Electronic Systems
Poznan 60-965
POLAND
Ph : +48 61 665 2836
Fax: +48 61 665 2840
tomasz.marciniak@put.poznan.pl

Treasurer:
Zdzisława Sobańska
211/215 Wolczanska Str., B-9 building
Lodz University of Technology
Lodz 90-924
POLAND
Ph : +48 42 631 2621
Fax: +48 42 636 2238
zdzislawa.sobanska@p.lodz.pl

Info Schedule Submitted by
Adam Dąbrowski
+48 61 665 2831
adam.dabrowski@put.poznan.pl

Entered in database: 4/16/2012

Program summary

	THURSDAY 27.09.2012	FRIDAY 28.09.2012	SATURDAY 29.09.2012	
09:00		Plenary Lecture II prof. Andrzej Czyżewski	Plenary Lecture III prof. Władysław Skarbek	09:00
09:15				09:15
09:30				09:30
09:45		SESSION 4 Human-Computer Interaction	SESSION 7 Stereovision and Video Coding	09:45
10:00				10:00
10:15	OPENING			10:15
10:30	Plenary Lecture I prof. Arvid Lundervold			10:30
10:45				10:45
11:00		COFFEE	COFFEE	11:00
11:15	SESSION 1 Image Processing I			11:15
11:30		SESSION 5 Image Processing II and Biomedical Applications	SESSION 8 DSP, Hardware and Applications	11:30
11:45				11:45
12:00				12:00
12:15				12:15
12:30	COFFEE			12:30
12:45			CLOSING	12:45
13:00	SESSION 2 Audio Processing I	LUNCH	LUNCH	13:00
13:15				13:15
13:30				13:30
13:45				13:45
14:00	LUNCH			14:00
14:15		SESSION 6 Audio Processing III		14:15
14:30				14:30
14:45				14:45
15:00				15:00
15:15	Hardware-software platform for measurements and processing of signals from National Instruments Sponsor Presentation by Wojciech Sommer			15:15
15:30		COFFEE		15:30
15:45				15:45
16:00	SESSION 3 Audio Processing II	CITY TOUR		16:00
16:15				16:15
16:30	COFFEE			16:30
16:45				16:45
17:00	SESSION 3 Audio Processing II – cont.			17:00
17:15				17:15
17:30				17:30
17:45				17:45
18:00				18:00
18:15				18:15
18:30				18:30
18:45				18:45
19:00	BANQUET (WHITE FACTORY)			19:00
19:15				19:15
19:30				19:30
19:45				19:45
20:00		PUB (IRISH PUB)		20:00
20:15				20:15
20:30				20:30
20:45				20:45
21:00				21:00

All sessions will be held at the Institute of Electronics lecture hall 416b, "Lodex" Building B9, 211/215 Wólczańska Str., Łódź
Lunches will be served at the university cafeteria, 3a Politechniki ave.

Dear Conference Participants!

Welcome to the NTAV/SPA 2012 conference held in Lodz. This year's scientific event joins two successful series of conferences: the New Trends in Audio and Video (NTAV) conference - organized first by the Wrocław University of Technology in 1994 and then biannually by various major technical universities in Poland, and the IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) conference - first held in 1993 and then organized yearly since 1999 by the Poznań University of Technology. Credit for the initiation, sustained support and organization of the NTAV and SPA conferences goes to professors Andrzej Dobrucki and Adam Dąbrowski, respectively.

NTAV/SPA 2012 is the second time these two conferences are joined as one event, the first time being in 2008 at the Poznań University of Technology. The aim of the NTAV/SPA 2008 was to gather researchers interested in various signal processing studies (including audio, video and image analysis). Similarly, the aim of this year's event, held four years later, is to bring together colleagues from a broad range of research subjects related to audio, image and video processing, digital signal processing theory, techniques and various applications including human-computer interaction systems and medicine.

An excellent introduction to the conference's broad scope is given by the three invited lectures prepared by distinguished scientists from Norway and Poland. Professor Arvid Lundervold from the University of Bergen presents an important image diagnostic technique known as Functional Magnetic Resonance Imaging (fMRI) and highlights its applications in basic and clinical neurosciences. The lecture by professor Andrzej Czyżewski from the Gdańsk University of Technology is devoted to the dynamically developing field of human-computer interaction and its novel applications in education and aids for persons with disabilities. Finally, professor Władysław Skarbek from the Warsaw University of Technology addresses the appealing concept of literal programming, i.e. the methodology of programming in high level languages resembling natural language more than program code.

The 42 regular papers accepted for NTAV/SPA 2012 were reviewed by two (or in some cases three) reviewers - members of the Scientific Committee. The contributions have been grouped into the following thematic fields: signal processing theory and algorithms, audio processing and acoustics, image processing and analysis, and finally human-computer interaction. All the accepted submissions printed in the conference proceedings will be indexed in the IEEE Xplore database. Selected papers pointed out by the Scientific Committee, after suitable extension, are invited for a publication in peer reviewed journals. It is worth noting here that, thanks to the support of the Audio Engineering Society, the abstracts of the NTAV/SPA 2012 papers devoted to audio and acoustics have been printed in the recent special issue of the Archives of Acoustics, a magazine published by the Polish Academy of Sciences.

The NTAV/SPA 2012 Organizing Committee acknowledges the financial support of the Ministry of Science and Higher Education. In this respect, special thanks are also due to the Dean of the Faculty of Electrical, Electronic, Computer and Control Engineering prof. Sławomir Wiak and the Dean-Elect prof. Sławomir Hausman. The contribution of the Polish Section of the Audio Engineering Society and the Polish Association of Theoretical and Applied Electrotechnics to the organization activities of the conference is also highly appreciated. Finally, sincere thanks are extended to all conference participants who submitted their work to the NTAV/SPA 2012.

We do hope that a wide scope of this joint conference will offer a unique forum for stimulating interdisciplinary discussion on new trends of all aspects of signal, image and video processing techniques and their widening important applications.

We wish you an enjoyable stay in Łódź!

On behalf of the Organizing Committee
Paweł Strumiłło

Contents

Plenary Lecture

SESSION 1: Image Processing I

SESSION 2: Audio Processing I

SESSION 3: Audio Processing II

SESSION 4: Human-Computer Interaction

SESSION 5: Image Processing II and Biomedical Applications

SESSION 6: Audio Processing III

SESSION 7: Stereovision and Video Coding

SESSION 8: DSP, Hardware and Applications

PLENARY LECTURES

Functional MRI - Signal Processing Algorithms and Applications

Arvid Lundervold, MD, PhD
Professor in Medical Information Technology

Neuroinformatics and Image Analysis Laboratory
Department of Biomedicine, University of Bergen
Jonas Lies vei 91, 5009 Bergen
&
Department of Radiology, Haukeland University Hospital
5021 Bergen, Norway

Functional magnetic resonance imaging (fMRI) investigations are increasingly important for the in vivo study and modeling of integrative brain functions in health and disease, where sophisticated mathematical and statistical algorithms for fMRI signal processing and interpretation have come into play. Apart from neuroanatomical, neurophysiological, and neuropsychological competence, the progress in cognitive neuroscience and brain mapping is critically dependent on expertise from other disciplines - such as statistics, computer science, and electrical and electronic engineering dealing with signal processing, circuits and systems. During the last years, there has also been a trend towards funding of "open science" consortia, providing huge and well-curated image data repositories together with advanced software tools and processing pipelines to be used and further developed by the research community.

In my talk, I will try to give glimpses from the big picture of this exciting and fast progressing field of fMRI and brain mapping, with the following outline:

- The human brain and information processing, spatial and temporal scales
- The history and basic principles of BOLD fMRI
- Signal, noise, and preprocessing of fMRI recordings
- Statistical analysis of preprocessed fMRI data
  - Voxel-wise analysis with the general linear model (GLM)
  - Multi-voxel pattern analysis (MVPA)
  - Independent component analysis (spatial and temporal ICA)
- Applications of fMRI in basic and clinical neurosciences
  - Neuronal encoding of sound and tonotopic maps
  - Visual processing and object recognition
  - Resting state fMRI and complex networks in health and disease
- Conclusions and Perspectives

New Applications of Multimodal Human-Computer Interfaces

Andrzej Czyżewski
Gdansk University of Technology, Multimedia Systems Department
80-233 Gdansk, Poland, ul. Narutowicza 11/12
e-mail: ac@pg.gda.pl

Abstract — **Multimodal computer interfaces and examples of their applications to education software and for the disabled people are presented. The proposed interfaces include the interactive electronic whiteboard based on video image analysis, application for controlling computers with gestures and the audio interface for speech stretching for hearing impaired and stuttering people. Application of the eye-gaze tracking system to awareness evaluation is demonstrated. The proposed method assumes analysis of visual activity of patients remaining in vegetative state. The scent emitting multimodal computer interface is an important supplement of the polysensoric stimulation process, playing an essential role in education and therapy of children with developmental disorders. A new approach to diagnosing Parkinson's disease is shown. The progression of the disease can be measured by the UPDRS (Unified Parkinson Disease Rating Scale) scale which is used for evaluating motor and behavioral symptoms of Parkinson's disease, employing the multimodal interface called Virtual-Touchpad (VTP) to support medical diagnosis. The paper is concluded with some general remarks concerning the role of multimodal computer interfaces applied to learning, therapy and everyday usage of computerized devices.**

***Keywords* — multimodal interfaces, video processing, speech processing**

I. Introduction

The research project entitled "Elaboration of a series of multimodal interfaces and their implementation to educational, medical, security and industrial applications" is realized by the authors of this paper. In this project, multimodal interfaces for the application in the area of medicine and of education (therapy of children, disabled persons) are developed. The proposed solutions are based on the human interaction with the computer using all five senses. Regarding the video analysis, methods of controlling the computer with gestures are developed. More advanced video processing algorithms are employed to tasks related to the Virtual Whiteboard and to diagnostics of the Parkinson's disease. The sound modality is used for slowing down the speech for improving its perception, and for training of the auditory and visual lateralization. Moreover, the system of polysensory training, based on a visual-auditory-kinesthetic stimulation, is developed. An important innovation is a computer-controlled scent emitting system. Additionally, a method of assessing patients' awareness in the locked-in syndrome, based on tracking of the eyesight focal point and on analysis of brainwaves, is developed. All above solutions are presented in the following sections of this paper.

II. Virtual Whiteboard

The Virtual Whiteboard is a computer system allowing for emulating an electronic whiteboard using a multimedia projector, a screen, a computer and a camera connected to the USB port (Fig. 1). The multimedia projector is mounted under the ceiling. A camera can be attached directly to the multimedia projector or placed on a stand at a distance from the screen so that its field of view fits to the video frame. The user is situated between the projector and the screen. An application installed on the computer controls the mentioned components and recognizes dynamic gestures, i.e. constituted by motion trajectories, and static gestures, e.g. palm shapes. Apart from the basic functionality of the whiteboard, which is entering the content, the system enables the user to interact with objects, e.g. rotation, zooming in/out, cropping and shifting is possible. A course of events during the work with the system can be saved and recreated, preserving time dependencies. The Virtual Whiteboard works with multimedia presentation browsers, providing functionality of browsing slides and adding notes. The solution has been implemented in 20 primary and high schools located in the Polish Pomerania region and in Gdansk Science & Technology Park in the education zone for children, called EduPark.

The principle of its functioning lies in subtracting the camera video stream from the stream displayed by the multimedia projector. Gestures are recognized in the resulting processed stream. The utilized image processing algorithms

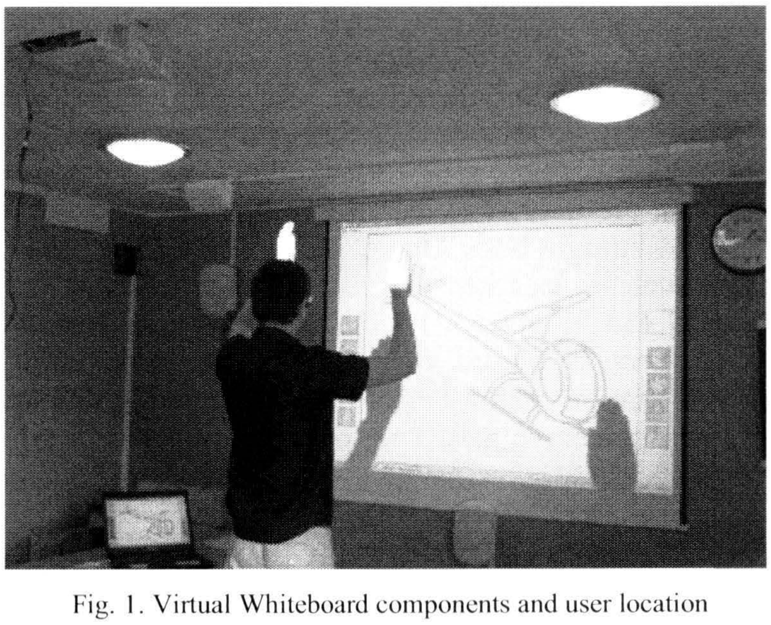


Fig. 1. Virtual Whiteboard components and user location

have been presented in the earlier paper [1]. For dynamic gesture recognition, an inference system based on fuzzy rules has been applied. Fuzzy rules have been used for modeling motion trajectories considering speed and direction of hand movement. Four fuzzy sets for directions: north, east, south, west, and four fuzzy sets for gesture speeds: very small, small, medium, high, are distinguished. The membership functions have triangular shape. Membership functions related to the performed motion associated with gesture classes was designed according to the Takagi-Sugeno zero order model. The output of the system is the result of the rule fired with the maximum membership degree. The decision threshold equal to 0.5 was utilized in the inference method. A value below this threshold means that the performed motion is associated with neither of the gesture classes. Kalman filters were utilized to smooth the motion trajectories, which enables to eliminate erroneous hand position detections. Static gestures are recognized using support vector machines of C-SVC type with RBF kernel. For the purpose of the palm shape parameterization, PGH (Pairwise Geometrical Histogram) histograms with 15 bins of height equal to 10 units were used. The method ensures insensitivity to hand rotation.

III. POLYSENSORY STIMULATOR

The polysensory stimulator is a tool dedicated to stimulating development of kinesthetic-perceptual functions in children with moderate or severe mental retardation. The aim of employing the system to the educational process is to help developing the visual and auditory perception, to increase duration of maintaining attention on stimuli, to improve visual-auditory-kinesthetic coordination, to develop and to enhance orientation in the body schema and space, to develop kinesthetic functions, and to boost language skills. The expected effect of the system application consists not only of development of particular functions but also of their mutual cooperation, i.e. perceptual-motor integration.

The system consists of a personal computer with the application installed, two monitors, a therapeutic mat, two USB cameras, 4 speakers of surround sound system, and a stand for the speakers and one of the cameras. Such a hardware platform has been also utilized during the development of the Multimedia System of Polysensory Integration [2], providing exercises for intellectually retarded pupils. The monitors are placed back to back, with one of them displaying the exercise screens to the pupil and the other displaying modified screens and controls to the therapist. The therapeutic area is designated by a stand construction with a square therapeutic mat lying in its center on the floor. The therapeutic mat has 9 square areas separated by straight lines. One of the cameras is placed on the floor in such a location that the person walking on the therapeutic mat is always visible in the image frame. The second camera is placed over the middle position of the therapeutic mat on such a height that the same requirement as for the floor camera is met. The speakers are positioned in the corners of the stand construction.

Interaction with the system is based on walking on the mat and thus choosing one of the squares at the time and bouncing to confirm the square choice. Occupying a particular square causes displaying an image or generating a sound associated with it. Bouncing produces image displaying and generating sound as well.

The system contains 11 exercises diversified in terms of difficulty level: simple ones for the intellectually developed children but challenging for the mentally retarded ones. The exercises involve various combinations of the task of searching for the image associated with the particular therapeutic mat square. By changing the location on the mat, the pupil changes the displayed image and the generated sound. The scheme of the exercises provided by the engineered software is fixed. A therapist can, however, customize the images and sounds to certain particular educational needs and age or interests of the pupil.

The research on the effectiveness of the system has been carried out for 8 weeks in the primary special school No. 26 in the Polish city Torun. Eight pupils, aged 8–17, took part in the therapy. Their high degree of motivation and interest in the equipment resulted in a relatively fast progress. It was especially noticeable in two spheres: spatial orientation & directions understanding and sense of causation, i.e. concentration of attention and understanding of cause and effect relationship while interacting with the system.

The system bases on the processed video streams obtained by subtracting the initial video streams that do not contain a pupil, from the video streams retrieved during the training. This principle regards both floor and overhead cameras. The subtraction method is based on the absolute difference. After subtraction, video streams are converted from RGB color space to gray scale and binary thresholded with a default threshold value. After the binarization, video streams are median filtered with a mask of size equal to 9 pixels (Fig. 2). The detection of pupil silhouettes is performed using a contour detection algorithm implemented in the OpenCV library.

IV. 3D HAND MODEL FOR SUPPORTING PARKINSON'S DISEASE OBJECTIVE DIAGNOSIS

Currently, more than 1% of the population aged over 60 suffers from the Parkinson's Disease (PD). Despite many conducted researches and various methods of treatment, the disease still remains incurable. The main problem in searching for new methods of treatment is lack of objective diagnosis method of this illness development. The disease if often rated in the so-called Unified Parkinson's Disease Rating Scale (UPDRS). During the examination, the patient is asked to perform a series of tests. Each test is then rated from 0 to 4, where 0 means that no symptoms were present and 4 that the patient was unable to perform the task. Usually, two different clinicians rate the same patient differently, therefore basing on

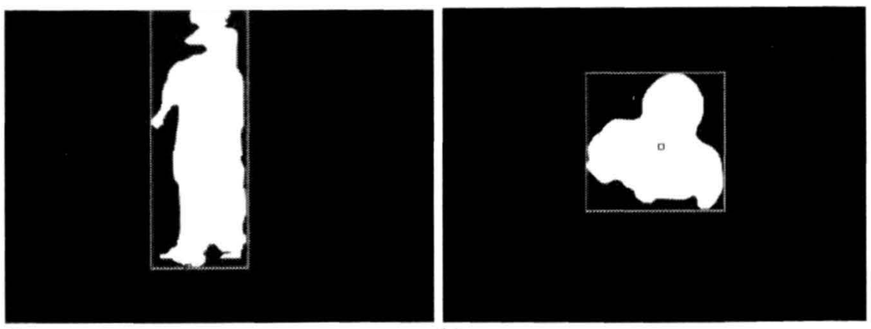

a) b)

Fig. 2. The processed image retrieved from the floor camera (a) and the overhead camera (b)

their diagnosis, thus it is impossible to assess deterioration or improvement in the patient's state. A development of new methods for objective diagnosis is strongly desirable.

The authors are conducting the research aiming to fill the existing gap. The developed algorithm utilizes the captured video stream to the rating of the performance objectify employing three UPDRS tests. These are: test 23 – finger tapping, test 24 – fist opening and test 25 – alternating hand movements. A web camera located on a special tripod captures the top view of the hand, which is placed on a black rubber pad. Due to such an approach, the separation of the object from the background is smoother. Most of the currently developed gesture recognition systems are based on the recognition of static key frames in the dynamic video stream. Such an approach would not guarantee, however, the needed information for diagnosing PD, as no information about the way the gesture performed is included. The algorithm presented in this paper solves this problem by visualizing the entire gesture on the motion curve (MC). First, a 3D hand model based on a single captured frame is created and then the reference animation of each gesture is rendered. The dynamic video stream is then compared with reference and the results are stored on MC [3]. The obtained MC provides a fundament for the objective classification. Additionally, in order to provide knowledge about naturalness of the movement, the classifier cooperates with the database of reference MCs generated from Motion Capture recordings. The entire process of the classification is depicted in Fig. 3.

In order to generate the reference animation, a single captured frame of patient's hand (neutral pose, fingers slightly spread) is converted into a binary image. The binarization threshold is dynamically computed, so the algorithm is becoming robust against changes in lighting. First, the image is split into R,G,B components and the histogram is analyzed for each channel. Basing on this analysis, a binarization threshold is computed. For a further computation, the image that separates the hand from the background in a best way is chosen. The objects in the image are searched with a chain approximation algorithm. The biggest object is the hand image. Then, a hand contour is created. The point cloud is inserted into the contour and the hand mask. The modified Delanuey triangulation algorithm is applied to generate a 2D mesh, which is then extruded into 3D. The bone structure is then added to the mesh. By modifying the bone structure, a reference animation is created.

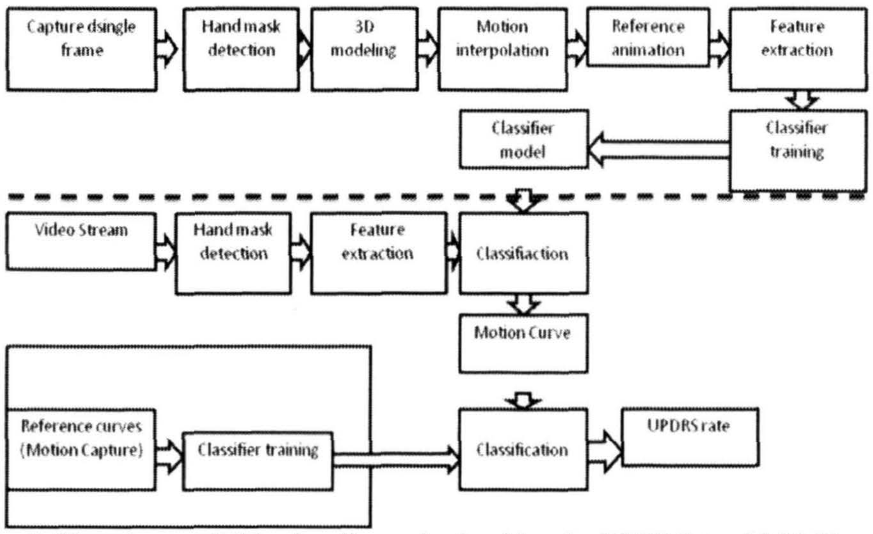


Fig. 3. The scheme of objective diagnosis algorithm for UPDRS test 23,24,25

A simple subtraction classifier is used for comparing the video stream to the animation. The MC is drawn. The maximum amplitude is interpreted as the midpoint of the performed gesture. An example of the MC is presented in Fig. 4. The presented methods allow for an easy distinction between motion of healthy people and PD patients. The frequency of extreme position crossing, smoothness of the MC and duration of the performing excursive provide objective parameters for the automatic classification.

V. Scent Emitting Multimodal Computer Interface

The aim of the project was to create a new kind of a computer interface, which could significantly enhance the polysensory stimulation process. There are numerous methods of treatment based on simultaneous inciting of different senses (e.g. Morning Circle or Multimedia System of Polysensory Integration [2]). The scent emitting computer interface should be able to diffuse an aroma in the classroom not only quickly, but also irrespectively of air humidity and temperature. To achieve this goal, it was decided to use a technique called Cold Air Diffusion. The advantage of this technique is very small dimension of aroma molecules: their size remaining below 1 micron (about 50 times smaller than molecules of deodorant). Small molecules allow for easier filling even large rooms with scent. In addition to that, small molecules easily combine with air molecules, thus the scent resides in the room for a long time.

Two versions of the scent emitting computer interface were developed in the framework of the project: the extended and the basic one. The extended version (Fig. 5) is able to emit four scents independently and it is equipped with the following sensors: gas, temperature, humidity and atmospheric pressure [2,4]. The basic version is able to diffuse one scent only and it has only the gas sensor built-in. However, it is possible to connect more than one device to the PC simultaneously. Both developed devices use USB or Bluetooth interfaces and they work with common PC computers.

The main use of the developed interface is extending functionalities of the polysensory stimulation. In this way, therapists will obtain a new tool that gives them opportunity to diffuse scents according to the children needs. Furthermore, it will be possible to correlate precisely the aromatherapy process with images, movies and sounds presented by the therapist. The interface can also be used as a vital complement of the multimedia educational applications. Emission of specific scents may cause e.g. encouraging inactive children or quieting hyperactive or restless children. It should be also pointed out that an appropriate choice of the aroma would make lectures on

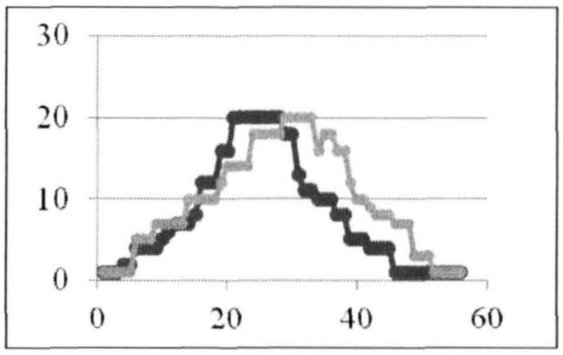


Fig. 4. An example result of comparison the MCs for healthy person (light line) and PD patient (dark line)

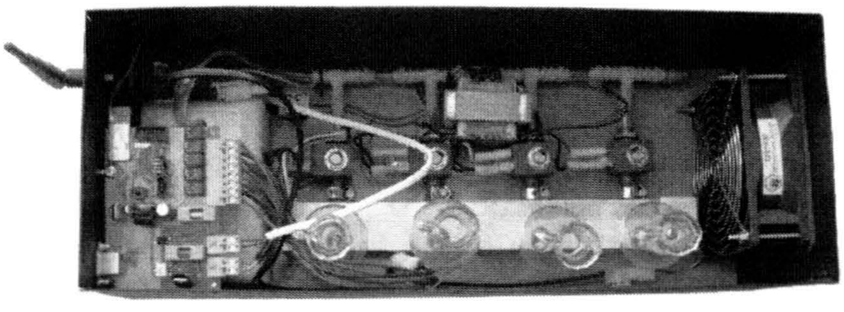


Fig. 5. Interior of the scent diffuser prototype (the extended version)

topics such as biology or materials science more attractive. The interface may be also used to diagnose neurodegenerative diseases in their early stages.

VI. System supporting communication between pupil and teacher

Problems with speech perception may cause an impairment in the language learning. Often, difficulties in speech perception are associated with peripheral auditory system disorders. In such a situation, the usage of a typical hearing aid (which works as a compressor) allows for eliminating this problem almost completely. However, not always speech perception problems are related to the peripheral auditory system. Chermak et al. estimated that 2% of children of the age between 6 and 10 years, suffer from the Central Auditory Processing Disorders (CAPD) manifesting with speech perception problems in complex acoustic conditions, difficulties in concentration, problems with perception of speech spoken with a high rate, problems with reading and speaking, etc.

One of the most known methods related to the problem of children with CAPD is the so-called FM system. It was designed in order to support speech perception during the school classes. The idea of it is simple: a teacher uses a wireless microphone located near his/her mouth; pupils listen to teacher's voice with their headphones. The audio signal between the devices is transmitted using the FM connection. This configuration allows one to reduce the reverberation and noise level related to classroom acoustics. The other group of similar systems are those that modify the time scale (TSM) of the input speech (time-expansion of the speech). As it was shown in the literature, in case of listeners with CAPD, time-expanding of the speech improves its perception.

Based on the methods presented above, a system which stretches (in real-time) teacher's speech and reproduces it on the pupils' headphones was designed. The system consists of two main parts: the speech stretching device and the workstation for performing hearing tests. In Fig. 6 the schema of the proposed system is shown. At the beginning, every pupil has to perform a series of hearing tests (the examination is supervised by the specialist). The workstation is used for performing peripheral auditory system tests (air conduction tonal audiometer and loudness scaling test), and central auditory system tests: Dichotic Digits Test (DDT), Duration Pattern Sequence Test (DPST), Pitch Pattern Sequence Test (PPST), Random Gap Detection Test (RGDT). Children with detected CAPD may use the proposed device.

The designed mobile device stretches the speech signal in real-time. It was necessary to design methods allowing to perform this operation and to ensure as small as possible difference in the duration of the input (original speech) and the output (stretched) signal. In Fig. 7 the block diagram of the designed algorithm is presented. The proposed algorithm is based on the assumption that the input signal is redundant. The noise and the stutter are considered redundant parts of the signal. In order to detect these events, the voice activity detector (VAD) and the stutter detector were used. Redundant parts of signal are removed and they are not stretched by the TSM algorithm. Additionally, TSM is performed non-uniformly, i.e.: vowels (detected by the vowel detector VRD) are stretched using a higher value of the scale factor than the consonants; dependently on the estimated rate of speech (ROS), fast speech is stretched more than the slow one. The TSM is performed using the SOLA algorithm (Synchronous Overlap and Add). The presented algorithm was implemented to the mobile phone.

VII. Auditory-visual attention stimulator

The proposed method of lateralization formation training was designed using a dedicated software. There are several scenarios in which it could be used. In the main scenario, a typical PC with the headphones serves as the training tool. In the extended version, some additional hardware is required, i.e. 3D display or an eye-tracking system. In Fig. 8 the block diagram of the proposed approach is presented. The main idea is to perform parallel stimulation of the eyesight and the hearing sense using digital signal processing techniques. Modification of the visual and hearing stimuli is performed in order to focus perception of those senses by the appropriate hemisphere (by means of the lateralization profile).

Speech modification is performed in two separate domains: TSM and signal amplitude variation. The TSM is obtained using the non-uniform real-time speech stretching algorithm [5]. The main purpose of the TSM algorithm is to modify duration of different speech units using various time scaling factors, in real time. Vowels, consonants and pauses are analyzed as the most significant speech units. The recorded speech and the assigned text are used during the training. The whole wave file is manually labeled in order to determine time stamps (LRC formatted) of the words. Image modification is

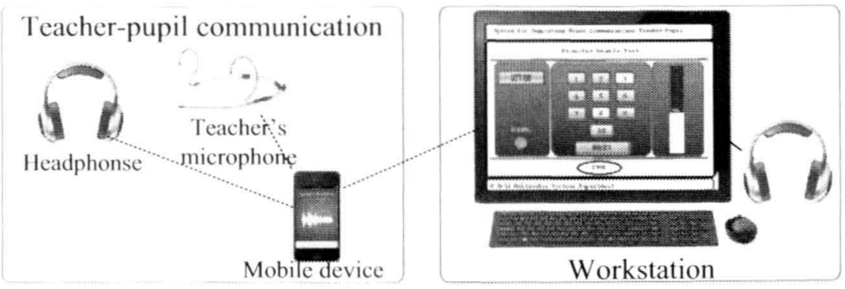


Fig. 6. Schema of the proposed system supporting communication between pupil and teacher

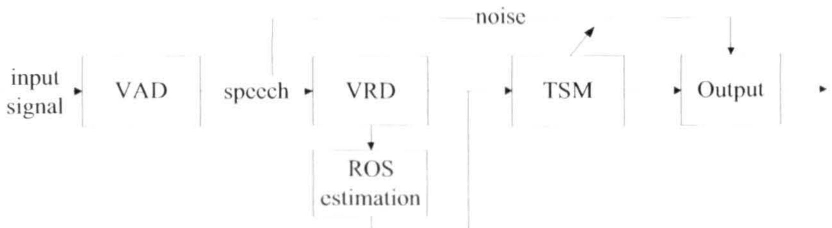


Fig. 7. Block diagram of the designed real-time speech stretching algorithm

fully synchronized with the speech signal. It can be achieved either at the level of words or sentences.

The developed system allows for visual stimulation of the eye using an appropriate adjustment of the following image parameters: color, brightness, contrast, size of the letters (including zoom profile - each letter is enlarged with different zoom factor). According to the classic Gestalt perceptual theory, the authors decided to reject the text highlight possibility. The user may not be able to distinguish the text from the highlighted background which becomes a figure.

As it is well known, reading with dyslectic fonts does not improve the reading speed. However, some specific types of reading errors are decreased. Therefore, in the presented system, it is possible to choose the appropriate font. In the extended version of the system, parameters associated with vision and hearing modification algorithms may be controlled by the eye-tracking system (Fig. 9). Such a system, developed in the Multimedia Systems Department was used for optimizing the training process [6,7].

VIII. CONSCIOUSNESS STUDY OF PATIENTS IN VEGETATIVE STATE

Evaluation of the consciousness level is the most important stage of rehabilitation of patients in vegetative state. The authors developed a consciousness test based on utilizing different multimodal interfaces: an eye-gaze tracking system and an EEG helmet. The research was conducted in the cooperation with therapists of the Therapeutic Care Center of "Light" Foundation in Torun, Poland.

First, hearing of each patient was examined, exploiting an objective method ABR (Auditory Brainstem Response). Estimation of the hearing threshold is reasonable because the therapists have to be confident that patients can hear their commands. It is worth mentioning that all 15 tested patients had normal hearing. The mean hearing threshold equaled to ca. 25 dB nHL. Then, a patient was subjected to tests utilizing the developed multimodal interfaces. Their interaction with the content displayed on the computer screen was studied using the eye-gaze tracking system. Also, the emotional state of each patient was examined during the polysensory stimulation, employing the EEG helmet [5]. When a therapist observed repeatable awareness symptoms, patient's electrical activity was analyzed in the context of intension of hand movement detection. A diagnosis of the consciousness state is related to comprehensive observation of the patient during several months. Therefore, evaluation of the consciousness level includes results obtained during sessions utilizing the eye-gaze tracking system and the EEG helmet. The scheme of the proposed method of consciousness state assessment is presented in Fig. 10.

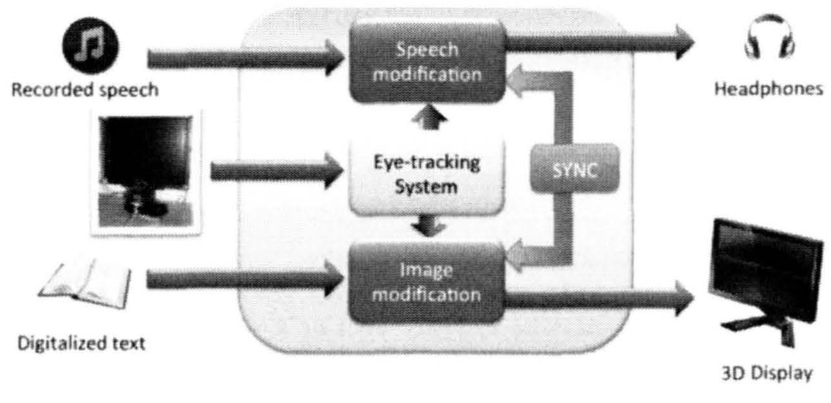
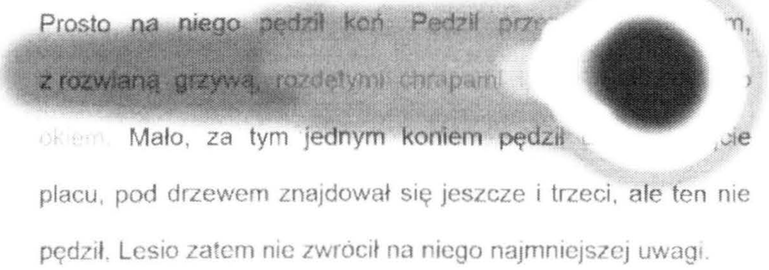


Fig. 8. Block diagram of the training system of the auditory-visual attention stimulator



Prosto na niego pędził koń. Pedził prz n,
z rozwianą grzywą, rozdętymi chrapami
okiem. Mało, za tym jednym koniem pędził cie
placu, pod drzewem znajdował się jeszcze i trzeci, ale ten nie
pędził, Lesio zatem nie zwrócił na niego najmniejszej uwagi.

Fig. 9. Sample text displayed on the screen - with feedback from the eye-tracking system

A. Eye-gaze tracking

Employing the eye-gaze tracking system in the consciousness study is an innovative approach in research conducted in this area. The authors and the therapists of the Therapeutic Care Center of "Light" Foundation carried out an initial experiment in order to assess the usefulness of the eye-gaze tracking system in patients' rehabilitation [7]. The confirmation of its usefulness enabled the authors to develop the so-called consciousness test including 12 exercises. Within this test, tracking of moving object, remembering, orientation in time and in place, understanding or logical thinking are studied. The ratio of correct tasks performed with the eye-gaze tracking system determines the evaluation of consciousness state. In case of several patients of the tested group, a repetitive value of correctly performed tasks ratio was observed during the six-month study period. The conducted experiment proved that the consciousness level in most of the tested patients was misdiagnosed.

B. EEG test

This section presents an attempt at determining the hand movement intention by analyzing the electric brain activity (EEG). The movement-related cortical potential (MRCP) could be noticed in the normalized EEG signal as the occurrence of the negative phase in the signal lasting from 0.5 to 2 seconds before the exact movement or its intention [1]. The detected MRCP is shown in Fig. 11.

Features obtained in multiple stages of processing the EEG

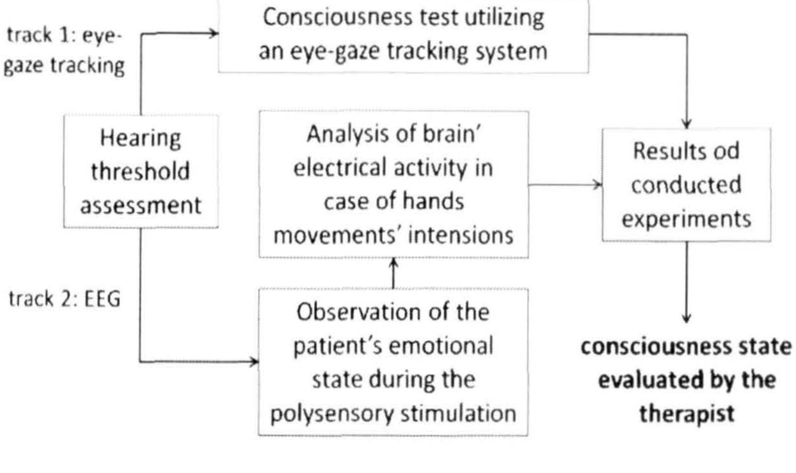

track 1: eye-gaze tracking
Consciousness test utilizing an eye-gaze tracking system
Hearing threshold assessment
Analysis of brain' electrical activity in case of hands movements' intensions
Results od conducted experiments
track 2: EEG
Observation of the patient's emotional state during the polysensory stimulation
consciousness state evaluated by the therapist

Fig. 10. The developed method of consciousness state evaluation

channels were the training input for the intelligent classifier. The abilities of Random Forest (RF) classifier were utilized. The experiment was conducted in the control group containing 15 healthy subject and 14 patients who were suspected to be in the locked-in syndrome. First, the data from movement execution for the control group were used to train the classifier and then testing was performed with the cross-validation method. The same experiment was then repeated for the movement intention. Next, data from the ME stage were used for training the classifier and the MI data were treated as the test set. Also, the opposite situation was examined. Then, both the ME and the MI models were tested with data obtained from patients' recordings. The obtained results proved that in all cases, the accuracy of right hand movement detection is 100%. The main reason for such a result is that during normal daily activities, activity of the left hemisphere of the brain (responsible for controlling the right side of the body) dominates, therefore additional changes in signal are simpler to notice. High efficiency in movement intention detection could also state about high rate of false positive values.

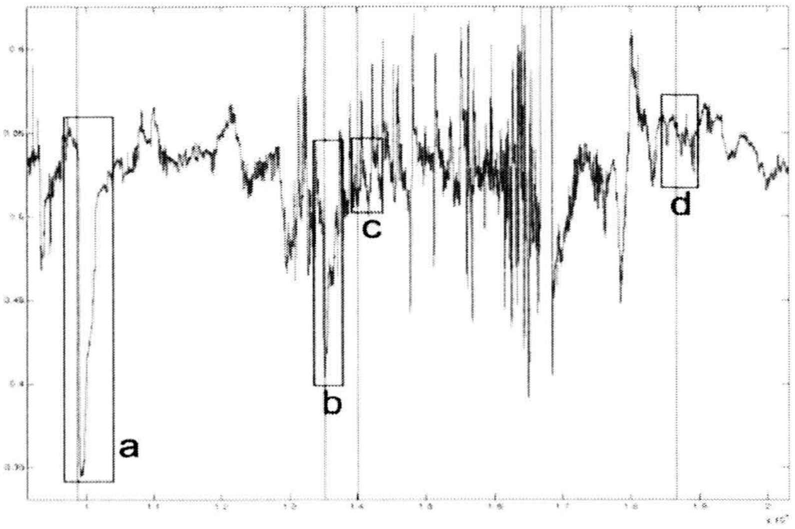


Fig. 11. Movement-related cortical potentials: a) b) MRCP, c) d) false positive value

IX. CONCLUSIONS

The developed multimodal computer interfaces were reviewed in this paper that enhance users' interaction with the computer by adding stimulation of senses that usually are not used in computer interfaces. Applying and in certain cases combining sound and video analysis with scent emission and brainwaves analysis constitutes an innovative aspect of the proposed solutions. A series of prototypes that were presented to end users during the tests, is a tangible effect of the described research.

It was proved that the proposed interfaces are positively received by the users participating in tests. The cooperation with schools and medical centers allows for assessing of the developed interfaces outside the research laboratory. The test results obtained from the end users prompted the authors to enhance the interfaces and to adapt them to the users' needs. In turn, constant technological developments encourage for further development of the proposed solutions. A wide range of possible end users prove that there is a need for this type of interfaces in many different areas. The scientific project "Elaboration of a series of multimodal interfaces and their implementation to educational, medical, security and industrial applications", realized in years 2008-2013, aims to commercialize the developed interfaces in the form of technology demonstrators, usable devices, publications and patents.

ACKNOWLEDGMENTS

Research funded within the project No. POIG.01.03.01-22-017/08, entitled "Elaboration of a series of multimodal interfaces and their implementation to educational, medical, security and industrial applications". The project is subsidized by the European regional development fund and by the Polish State budget.

The author of this review paper would like to give credits to his following co-workers for their valuable input to the described research: Katarzyna Kaszuba, Łukasz Kosikowski, Rober Kosikowski, Bożena Kostek, Bartosz Kunka, Adam Kupryjanow, Michał Lech, Piotr Odya & Piotr Suchomski.

REFERENCES

[1] M. Lech, B. Kostek, "Hand gesture recognition supported by fuzzy rules and Kalman filters," Int. Journal of Intelligent Information and Database Systems, vol. 6, No 5, pp. 16, 2012.

[2] M. Lech, B. Kostek, "Human-computer interaction approach applied to the multimedia system of polysensory integration," in New Directions in Intelligent Interactive Multimedia Systems and Services, Studies in Computational Intelligence, vol. 226/2009, W. Damiani, Ed, Berlin: Springer-Verlag, 2009, pp. 265 – 274.

[3] K. Kaszuba, B. Kostek, "A new approach for an automatic assessment of a neurological condition employing hand gesture classification," Multimedia/Multimodal Human-Computer Interaction in Knowledge-based Environments of the Intelligent Decisions Technologies Journal, L.C.Jain, G. Phillips-Wren, R.J. Howlett, Eds, IOS Press, 2011.

[4] A. Czyżewski, P. Odya, J. Smulko, G. Lentka, B. Kostek, M. Kotarski, "Scent emitting multimodal computer interface for learning enhancement," MIMIC 2010 – Fourth International Workshop on Management and Interaction with Multimodal Information Content, Bilbao, Spain, 2010, pp. 142 - 146.

[5] A. Kupryjanow, A. Czyzewski, "Improved method for real-time speech stretching," Intelligent Systems Technologies, vol. 6, pp. 1-9, 2012.

[6] L. Kosikowski, A. Czyzewski, "Binocular vision impairments therapy supported by contactless eye-gaze tracking system," Lecture Notes in Computer Science, vol. 6180, pp. 373-376, 2010.

[7] B. Kunka, A. Czyżewski, A. Kwiatkowska, "Awareness evaluation of patients in vegetative state employing eye-gaze tracking system," International Journal on Artificial Intelligence Tools, vol. 21, no. 2, pp. 1240007-1–11, 2012.

[8] K. Kaszuba, B. Kostek, "A bimodal approach to brain-computer interaction measurements," Signal Processing Algorithms, Architectures, Arrangments and Application Conference, IEEE Signal Processing Chapter, Poznań, Poland, 2011, pp. 1-6.

[9] A.T. Boye, U.Q. Kristiansen, M.Billinger, O. Flex do Nascimento, D. Fatina, "Identification of movement-related cortical potentials with optimized spatial filtering and principal component analisys," Biomedical Signal Processing and Control, vol.3 , pp.300-304, 200

Integrated Literate Programming for Structured Light Camera Calibration

Władysław Skarbek
Faculty of Electronics and Information Technology
Warsaw University of Technology

ABSTRACT — Literate programming is a concept initially conceived by Professor Donald Knuth who had applied it in practice while developing his TEX formatting system. In my integrated approach I advocate the concept of literate programming in a single book, i.e. all levels of a created application (including the descriptions of new algorithms) are joined together with all source codes in one PDF document, which is created from a tree of Latex documents. The code granularity can be of any detail. Even a single line of code is possible, surrounded by necessary comments, if needed. Codes can be composed sequentially and hierarchically. The presented paper will discuss advantages of the proposed programming methodology for developing software prototypes of applications within digital media one the example of calibration algorithms for a structured light camera system.

INDEX TERMS — literate programming, software documentation, structured light method, camera calibration

I. INTRODUCTION

The pioneer of literate programming is Donald Knuth – the genius of Computer Science. In 1980's, he incorporated this concept in his WEB language (an extension of Pascal) while implementing the TEX formatting system. Actually, *The TEXbook* of Knuth had been written using WEB [12]. There is also CWEB – the C version of WEB. Let us listen to words of wisdom expressed by Knuth: *The structure of a software program may be thought of as a "WEB" that is made up of many interconnected pieces. To document such a program we want to explain each individual part of the web and how it relates to its neighbors. The typographic tools provided by TEX give us an opportunity to explain the local structure of each part by making that structure visible, and the programming tools provided by languages like C make it possible for us to specify the algorithms formally and unambiguously. By combining the two, we can develop a style of programming that maximizes our ability to perceive the structure of a complex piece of software, and at the same time the documented programs can be mechanically translated into a working software system that matches the documentation.*

Another excerpt is: *Literate programming is a methodology that combines a programming language with a documentation language, thereby making programs more robust, more portable, more easily maintained, and arguably more fun to write than programs that are written only in a high-level language. The main idea is to treat a program as a piece of literature, addressed to human beings rather than to a computer. The program is also viewed as a hypertext document, rather like the World Wide Web.*

More on Knuth's original approach to literate programming can be found at the web site [14].

It seems that there were only few followers of Knuth. One positive example of the technical text written in literate programming style is the book on computer graphics [4].

Just after encountering this book, I decided (in September 2009) to write the Integra application to join the literate programming concept to one-text-many-views concept which was already used by me for combining presentation slides with lecture notes. Actually, the preprocessor for the latter approach was written by the research assistant Marcin Morgos in 2003, according to my requirements.

I implemented Integra in Python and the program serves me well for flexible development of lecture notes where Python and C/C++ code is extensively used for teaching multimedia algorithms.

In my integrated approach I advocate to the concept "system in a single book", i.e. all levels of an application being created (including description of new algorithms) are joined together with codes

This is the paper accompanying the invited lecture to NTAV/SPA 2012.

in one PDF document, which is created from a tree of Latex documents. The code granularity can be of any detail. Even the single code line is possible, surrounded by necessary explanations, if needed. Codes can be composed sequentially and hierarchically.

In this paper I would like to share my experience how the literate programming works when the algorithms are developed or tuned while a new application scc is designed and implemented in usual conditions of research uncertainty.

The application scc refers to calibration of structured light camera. Actually, our camera consists of three cooperating components: high speed infrared projector (200 fps), high speed infrared camera (200 fps, 640x480), and HD camera (25 fps). It is supposed that the calibration data is to be used by a 3D viewer application, working in real time when the system observes an arbitrary 3D scene.

Besides implementing and integrating of calibration algorithms the scc goal is to provide timed information for its users on

1) actions to be performed in the given stage of calibration,
2) data items to be updated in application configuration files,
3) values of calibration parameters which have been established,
4) quality measures for particular calibration steps.

There are four channels for user interaction:

1) log window with prompts and messages,
2) pop-up menus for selection of relevant steps,
3) interpreter console to show results,
4) configuration files to tune calibration setup parameters.

There are two modes of scc activity:

1) simulation mode: stripes projector, stripes camera, and video camera are simulated in virtual environment using OpenGL 3D graphics standard,
2) real mode: previewers windows of simulated cameras are replaced by previewers windows of real cameras.

II. Introduction to Integrated Literate Programming

The literary programming is a philosophy of software creation, which follows from a belief that computer programs can be written by programmers similarly to literary texts which are always addressed to potential readers. In this case there are two groups of readers:

- computers reading the created source code to execute it,
- programmers (including the primary software author) interested in code maintenance and developing, focusing rather on the documentation of the available software.

The *Integrated Literate Programming* is an attempt to integrate in one Latex document or in a tree of Latex documents content for both addressees.

Within *the writer's metaphor* the software creator should like a true writer enhance his/her literary style in order for both the computer and the man to benefit from reading his/her work.

1) Benefits for computers:
   a) executing programs correctly,
   b) using computing resources effectively.
2) Benefits for designers and programmers (usually the same persons):
   a) understanding of design requirements and decisions;
   b) understanding of source code, i.e. he/she can say: I can test the code, I can change it, and I can develop it.

In the above philosophy of software build, the source code and its documentation are equally important components. Fragments of documentation and source code are interleaved like in novels, scene and person description with monologue or dialog.

Let us observe that a change of application requirements means the change of documentation together with the relevant source code – like a new person entrance means its participation in the next dialog.

At the end, a document emerges in the form of one or more Latex (textual) files with optional inclusion of image, video, and audio materials. From such manual, besides a comprehensive system manual in PDF format, the source code files are extracted and optionally multimedia presentation or system web pages could be generated, too.

A. *Integra directives – requirements*

Integra is controlled by simple directives which provide tools:

1) to specify values for attributes of integration process,
2) to define textual fragments,
3) to compose textual fragments.

The basic requirements for `Integra` commands include:

1) Commands are embedded into Latex document within Latex comments.
2) Types of commands:
   a) specifying of `Integra` outputs,
   b) embracing textual fragments,
   c) assigning fragments to outputs,
   d) assigning names to fragments,
   e) fragments nesting,
   f) fragments concatenation,
   g) defining textual blocks.
3) *Authors expectations:* directives should be intuitive for understanding and compact for writing.
4) *Machine expectation:* directives should be quickly interpreted by Python.

B. Important directives

1) Assignments in keyword notation: `<attribute name> = <attribute value>`
2) Lists of attribute assignments are included in the single line of Latex comment.
3) Selected directive lines begin from:
   a) `%i%` if it identifies the output document,
   b) `%(%` if it starts a textual fragment,
   c) `%)%` if it ends a textual fragment,
   d) `%n%` if it stands for a textual fragment to be nested.

Example:

```
%i%id=4,file='hash.py'
...
%(%into=4
...
   %n% name of nesting fragment
...
%)%
```

C. A simple example

Suppose that within a *Software Engineering* course a problem of saving arbitrary precision integer numbers is considered.

BEGIN of EXAMPLE

Let us consider, for instance numbers produced by a factorial function. For the argument `n` the factorial value $n!$ is to be saved to the file `fname` by the function `factorialToFile`

The Python source code file is `sf.py` established in default directory of sources.

Implementation is based on standard arithmetic operations and functionalities of `file` class.

```
3  def factorialToFile(n,fname):
4      s = 1
5      for i in xrange(2,n): s *= i
6      text = str(s)
7      fp = open(fname,'w')

9      #< Formatting while saving to file >#
14     fp.close()
```

The result is subdivided into lines of 60 digit each:

```
9  ls = len(text)
10 for k in xrange(0,ls,60):
11     lp = min(ls,k+60)
12     fp.write(text[k:lp]+'\n')
```

In the source file, the implemented function precedes its conditional testing (e.g. by the way: fix character coding in the first line of the module):

```
1  # -*- coding: utf8 -*-

3  #< Computing factorial and saving to file >#
15 if __name__=='__main__':
16     n = 1000; fname = ''+str(n)+'.txt'
17     print '%d! to file %s ...' % (n,fname)
18     factorialToFile(n,fname)
```

The first four lines (out of 43) from the file `1000.txt` are given below:

```
402387260077093773543702433923003985719374864210714632543799
910429938512398629020592044208486969404800479988610197196058
631666872994808558901323829669944590997424504087073759918823
627727188732519779505950995276120874975462497043601418278094
```

END of EXAMPLE

Note, that while "the system manual" includes the text between the line *BEGIN of EXAMPLE* and the line *END of EXAMPLE*, the actual content of integrated file `sf.py` corresponds to line numbers in the manual:

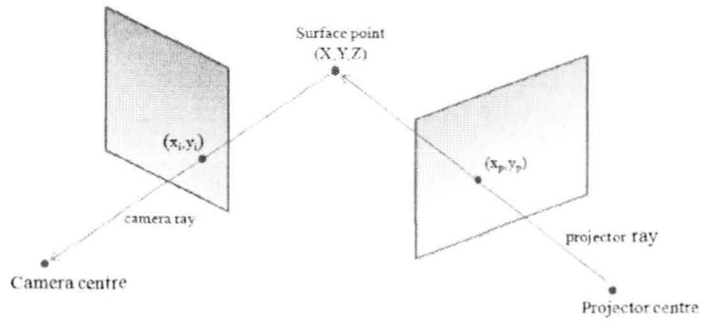
```
#  *- coding: utf8  *
def factorialToFile(n,fname):
    s = 1
    for i in xrange(2,n): s *  i
    text   str(s)
    fp   open(fname,'w')
    ls = len(text)
    for k in xrange(0,ls,60):
        lp   min(ls,k+60)
        fp.write(text[k:lp]+'\n')
    fp.close()
if __name__    '  main  ':
    n   1000; fname   ''+str(n)+'.txt'
    print '%d! to file %s ...' % (n,fname)
    factorialToFile(n,fname)
```

III. Introduction to Structured Light Camera Calibration

Structured light is a method of light volumes forming which is useful in surface modeling when projected on them [1], [2], [3], [5], [6], [7], [8], [9], [10], [11], [13], [15].

Stripe images are a mean of 3D space subdivision into sectors to which a unique code can be assigned. The code can be detected (read out) in stripe camera image which acts with the same speed as the projector while acquiring image of the illuminated scene in the spectrum range.

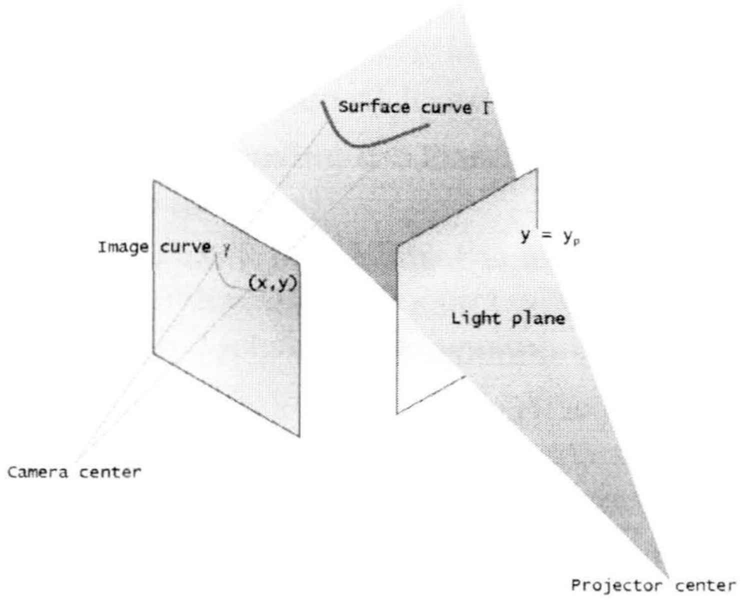
A. Beginning of light structuring: single ray



A 3D point registration:

- *Single ray design:* set a single pixel (x_p, y_p) in the graphics card frame buffer with adequate RGB values.
- *Project the ray on a surface:* ensure that the reflected ray from an unknown point (X, Y, Z) differs in colour wrt to its spatial neighbours.
- *Get an image of the surface:* make a camera single shot.
- *Detect image (x_i, y_i) of (X, Y, Z)* : apply colour segmentation procedure.
- *Compute an approximation $(\tilde{X}, \tilde{Y}, \tilde{Z})$ of the point (X, Y, Z)* : apply the Least Square Method to find a point of equal to minimal distance from projector's ray (x_p, y_p) and camera's pixel ray (x_i, y_i).

B. Plane of light rays

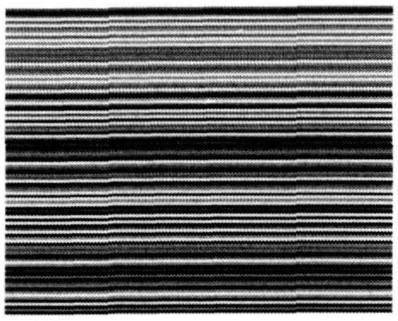


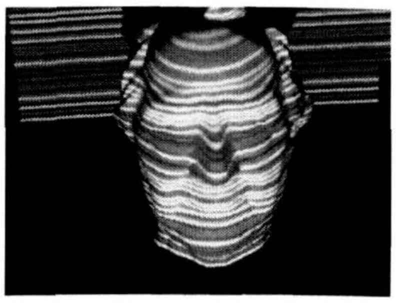
- Plane of rays ⟶ curve Γ on object's surface:
  1) *Emission of light plane $y = y_p$.*
  2) *Image registration* for modeled object.
  3) *Detection of image curve* γ = the image of Γ.
  4) *Approximation of spatial curve* Γ.

C. A systemic view

- 2D imaging as 2D system:
  1) *Input:* the plane of light rays $y = y_p$.
  2) *State:* the unknown curve Γ on object's surface.
  3) *Output:* the discrete curve γ.
- 3D Modeling as the identification problem:
  1) *Given:* image(s) of 3D object illuminated by structured light.
  2) *Identification of system output:* the detection of discrete curve γ in camera's image.
  3) *Identification of system output:* the detection of light plane $y = y_p$. generating γ.
  4) *Identification of system state:* the approximation of the curve Γ on the object's surface.

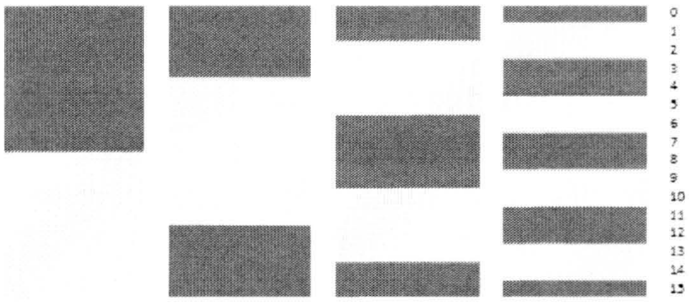
D. Advancing light structuring – stripes



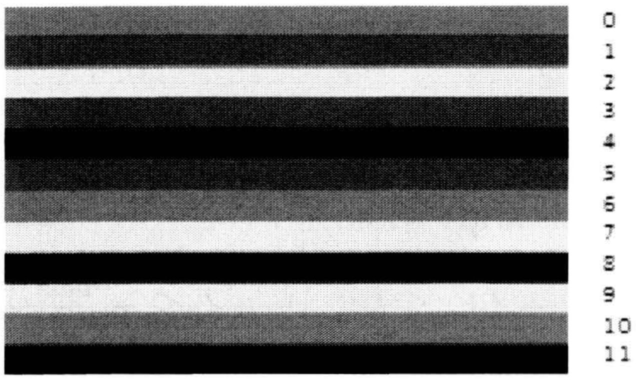


- *Stripe:* colour (in RGB), orientation (H or V), location (first line index and width)
- *Stripe pattern* is the list of disjoint stripes of the same orientation, covering the projector's image frame.

What is important, the stripe sector is a pyramid with its apex located within the volume of the stripe projector and with unspecified base. In the stripe camera, for each pixel the sector's code is readout for the stripe sector with the view pyramid determined by the pixel. Such readout is possible if in the intersection of both pyramids there is a surface reflecting projected light. Finally, the combination of pixel index and sector index determines the depth of the surface.



- Two types of codes: sequential and parallel in display time.
- *n-sequential colour code of a pixel* wrt a sequence of n stripe patterns of the same orientation: the sequence of n colour values of the stripes, the pixel belongs to.
- *n-sequential stripe code* is a sequence of n stripe patterns with the same orientation, such that the pixels with the same n-sequential colour code create a connected set of pixels.



- *n-stripes neighbourhood of a pixel* wrt a stripe pattern is a stripe the pixel belongs to and the next $n-1$ stripes (in axis order with $\bmod$ warping for the trailing stripes).
- *n-parallel colour code* of a pixel wrt a stripe pattern is the sequence of colors for its n-stripes neighbourhood.
- *n-parallel stripe code* is a stripe pattern such that the pixels with the same n-parallel colour code create a connected set of pixels.

E. Gray patterns

- Gray code of length n over alphabet $\Sigma_s \doteq \{0, \ldots, s-1\}$ is a one-to-one mapping $G_{s,n} : [0, s^n) \longrightarrow \Sigma_s^n$ for which $G_{s,n}(i)$ and $G_{s,n}(i + 1 \bmod s^n)$ differ on exactly one position.
- Domain of $G_{s,n}$ is the set of indexes for items to be encoded.
- Item to be encoded – for instance a stripe of light.
- Gray code of length one:
  $\Rightarrow$ The code $G_{s,1}(i) \doteq i$, $i \in \Sigma_s$ is Gray code of length one.
- Inverse Gray code: $\Rightarrow$ Let $G'_{s,n}$ be an *inverse Gray code* defined wrt to Gray code $G_{s,n}$ as follows:

$$G'_{s,n}(i) \doteq G_{s,n}(s^n - 1 - i), \ i \in [0, s^n)$$

  Then $G'_{s,n}$ is the Gray code over the alphabet Σ_s.
- Binary Gray codes in 2D tables:
  - item index is array column index;
  - code position is array row index.
- Example for $s = 2, n = 1$:

$G_{2,1}$	[0]	[1]
[0]	0	1

- Example for $s = 2, n = 3$:

$G_{2,3}$	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]
[0]	0	0	0	0	1	1	1	1
[1]	0	0	1	1	1	1	0	0
[2]	0	1	1	0	0	1	1	0

F. Departure from Gray codes for calibration and modelling

In the next section we derive another code to define stripes with better properties than Gray code based patterns.

For our camera, pixel dependent thresholding is more important for accurate camera image binarisation. The repetition factor $r \in [1, 3]$ will characterize the admissible code.

Another property is a balance between white and black areas in the stripe pattern image. This is achieved by symmetrical offset in the table of admissible codes. This leads us to valid codes used for calibration and modeling.

There are two correspondences as the results of the structured light camera calibration:

1) (stripe camera pixel, valid code index) $\mapsto$ depth index,
2) (stripe camera pixel, valid code index) $\mapsto$ video camera pixel.

Since optical distortions of cameras contribute with high share, and projector with less impact, both correspondences are highly nonlinear. Therefore, the number of modeling parameters of both correspondences is intractable. Hence they are approximated by interpolation process using experimentally established correspondence tables.

IV. Case Study: Stripes Module

As the implementation of the whole `scc` application is too large to be shown here I have selected the `stripes` module to present literary programming in details. It is more image processing oriented than other modules and it fits more to the conference scope.

The `stripes` module handles all aspects of stripe images handling including index images creation and corner detection from them.

A. *Stripe patterns – code image approach*

Any sequence of k binary images defines a *code image* including at most 2^k different binary codes. Since we want to use such coding to subdivide the 3D space into identical parts, the corresponding areas in code image, so called level sets, should be identical too. The simplest areas of this kind are horizontal or vertical stripes of equal height L_h or width L_w. For structured light camera calibration, we need more fine 3D space subdivision and then square areas obtained as intersections of horizontal and vertical stripes will fit.

High speed stripe projectors require saving all stripe images in its memory. Then its capacity determines the upper bound k_{max} for total number of horizontal k_h and vertical k_v stripe images:

$$k_h + k_v \leq k_{max}$$

This is the limit used in the calibration application and for the off-line 3D modeling of still objects. In practice $k_h, k_v < 10$, and $k_{max} > 1000$. Therefore the above condition is satisfied in contemporary technology state of art.

In 3D player we use only one-directional stripes, and then the speed $f_{ps}^{(s)}$ of the stripe projector and stripe camera and the speed $f_{ps}^{(v)}$ of the video (texture) camera, measured in actually registered image frames per second, matter:

$$k \leq \frac{f_{ps}^{(s)}}{f_{ps}^{(v)}}$$

where $k = k_h$ or $k = k_v$.

Obviously for the projector with resolution $W \times H$ the further constraints for k_h and k_v :

$$k_h \geq \log_2 H, \quad k_v \geq \log_2 W$$

Here, we have assumed for the finest stripes of one pixel width. However, in practice, the registration of borderline pixels is error prone. Since both sides of border can be affected, the borderline could be wider than one pixel. Therefore, perfect codes are expected at ratio at least $100(L-2)/L\%$. For instance for stripe width $L = 4, 10, 20$ we get $50\%, 89\%, 90\%$, respectively.

For system calibration we need higher ratio of code recognition and single line resolution of code image. By trading effective code image resolution with calibration time we display the same stripe image L times. At i-th time each image is displayed at offset i rows (in case of horizontal stripes) and at offset i columns (in case of vertical stripes).

Note that the code index $j + i/L$ for pixel with wrong code at i-th offset if the perfect code index j at offsets $i' = 0, \ldots, i-1$, is found and the perfect code index $j+1$ is detected at offsets $i' = i + 1, \ldots, L-2$.

Since projector's memory is large enough, we can keep all original images and their offset images together for stripe width L :

$$Lk_h + Lk_v \leq k_{max} \longrightarrow L \leq \frac{k_{max}}{k_h + k_v}$$

For $k_v \leq k_h = 10, k_{max} = 2000$ we have $L \leq 100$. With $f_{ps}^{(s)} = 200$ we get not more than 10 seconds to collect all calibration data at the fixed relative position of the video camera and projector screen. Therefore even for such extreme case mechanical movements of the devices determine the whole time of calibration procedure.

In particular case if $W = 640 = 5 \cdot 2^7, H = 480 = 3 \cdot 5 \cdot 2^5$ then $L = 5, 10, 20$ are feasible and the corresponding number of horizontal stripes $K_h = H/L = 96, 48, 24$ and the number of horizontal stripes $K_v = W/L = 128, 64, 32$.

Since each stripe has a unique code, the number of codes $C_h \geq K_h$ and $C_v \geq K_v$. The number of codes C_h (C_v) will be determined by the construction of code table which in turn will depend on number of bits k_h (k_v), and the zero-one repetition factor r. For instance $r = 2$ means that in each bitstring of code table we have at least two zeros and least two ones.

1) Code table construction: Let us follow the following reasoning:

1) The light emitted from the projector is always an additional light in the 3D scene.
2) Therefore we cannot distinguish all zeros from all ones.
3) Repetibility condition on the level r : qantity of zeros $\geq r$, quantity of ones $\geq r$.
4) For $k = 4$ stripe images the code table (codebook) consists of the entries: $0011, 0101, 0110, 1001, 1010, 1100$.
5) In general for k, r we get the number C_{kr} of codes:
   a) If $r = 1$ then $C_{k1} = 2^k - 2$,
   b) If $r = 2$ then $C_{k2} = 2^k - 2k - 2$,
   c) If $r = 3$ then $C_{k3} = 2^k - k(k + 1) - 2$:
   $$C_{k3} = 2^k - 2 - 2k - \binom{k}{2} - \binom{k}{k-2} = 2^k - 2 - k - k^2$$
6) For $r = 2$, one of the reasonable binarisation procedure is defined by thresholding with regard to $T(i, j)$:
   $$T(i, j) = \frac{g_{l_1}(i,j) + g_{l_{k\ 2}}(i,j)}{2},$$
   $$g_{l_0}(i, j) \leq g_{l_1}(i, j) \leq \cdots \leq g_{l_{k\ 1}}(i, j)$$
   where $g_l(i, j)$ jest l-th gray scale value in sorted k values at pixel (i, j)
7) Notation `argsort`:
   $$[l_0, l_1, \ldots, l_{k-2}, l_{k-1}] = \text{argsort}[g_0, g_1, \ldots, g_{k-2}, g_{k-1}]$$

We begin from importing mathematical tools

```
# -*- coding: utf8 -*-
import numpy as np
```

image file reading and writing functions

```
from image.imageIO import imageFileToMatrix
from image.imageIO import matrixToImageFile
```

and image filtering function

```
from mfilter.mfilter import\
    __filter2d__ as f2d
```

The module is built of class `Stripes`, and of the test part implemented by the function `testStripes`.

The class `Stripes` collects all functionalities we connect to stripe image construction, processing, and analysis.

B. Structure of `stripes.py`

1) Class `Stripes`:

```
class Stripes:
        #< Stripe constructor >#
        #< Code admissibility >#
        #< Code table size >#
        #< Number of stripe images >#
        #< Stripe image generator >#
        #< Stripe images from file >#
        #< Index image built >#
        #< Corner extractor >#
```

2) Testing function:

```
def testStripes():
        #< Test generating of stripes >#
        #< Test generating of index image >#
        #< Test corner extractor >#
if __name__=='__main__': testStripes()
```

C. `Stripes` *constructor*

It is reasonable to make the stripe generating independent of the projector'se resolution. To achieve this goal a single bit is assigned for each stripe. Then the color pixels in stripe are obtained by scaling and copying the bit according the required resolution.

Therefore the constructor of object `Stripes` defines the list of valid codes tables – one table per stripe image. In order to understand the process of generating valid codes follow the definitions:

1) `k` – codeword length,
2) `K max` – the number of all possible codewords of length k : $K_{max} = 2^k$,
3) `r` – zero-one repetition factor ($2r < k$)
4) admissible codeword – it consists of at least r zeros and r ones,
5) `C_kr` – the number of admissible codes of length k, with repetition factor r,
6) valid codeword – the admissible codeword with index between the `first` (f) and the `last` (l) in the sequence of admissible codes,

7) K – the number of valid codes:

$$K = l - f + 1,\quad K \le C_{kr}$$

8) Symmetric range of valid codes we get if:

$$f \doteq \left\lfloor \frac{C_{kr} - K}{2} \right\rfloor,\quad l \doteq f + K - 1$$

In our approach K is a compromise between projector image resolution and stripe width. Having K we find the first and the last valid codes from the above condition on symmetric range.

1) Constructor header:
   a) start of method header:

```
def __init__(
    self,
```

   b) codeword length:

```
    k,
```

   c) number of valid codes:

```
    K,
```

   d) code repetition factor:

```
    r=2,
```

   e) end of method header:

```
):
```

2) Save all parameters for future use:

```
    self.k = k; self.K = K; self.r = r
    self.th = 1
```

3) Prepare codes – the list of byte tables of size K each to store bits of codewords. It is column-wise representation of conceptual code table: code[i][j]==1 if and only if i-th bit of j-th codeword is equal to one. In other words: the first byte table represents the least significant bit of all codewords, the last byte table represents the most significant bit of all codewords.

```
    self.codes = []
    for i in range(k):
        self.codes.append(
      np.zeros(self.K,dtype=np.uint8)
      )
```

4) The valid codes are obtained from admissible codes by skipping an equal number of leading and trailing admissible codes:

```
    C_kr = self.countCodesNumber()
    if (self.K>C_kr):
        et = "Too␣many␣stripes!"
        raise Exception(et)
    first = (C_kr-K)/2
    last = first+K-1
```

5) Prepare a lookup table from arbitrary codes on k bits to their indexes in the valid code table – 0 stands for invalid code:

```
    K_max = 2**k
    self.code_index =\
        np.zeros(K_max,dtype=np.int32)

    #< Fill index and code tables >#
```

6) Create index and code tables:

```
j_a = 0 # index of valid code
j_b = 0 # index of admissible code
for j in range(0,K_max):
    if not self.isAdmissible(j):
        continue
    if first<=j_b:
        code = j
        self.code_index[j]=j_a+1
        for i in range(k):
            if code&1:
                self.codes[i][j_a]=1
            code >>= 1
        j_a += 1
    j_b += 1
    if j_b>last: break
```

D. Admissible codes

1) Counting ones and zeros for the codeword code:

```
def isAdmissible(self,code):
    n_zeros = 0; n_ones = 0
    for i in range(self.k):
        if code&1: n_ones += 1
        else: n_zeros += 1
        code >>= 1
    if (n_zeros>=self.r) and\
      (n_ones>=self.r):
        return True
    else:
        return False
```

2) Counting admissible codes:

```
@staticmethod
def countCodes(k,r):
    K_max = 2**k
    if r==1: return K_max-2
    elif r==2: return K_max-2-2*k
    elif r==3: return K_max-2-k-k*k
    else:
            et = "Repeat␣factor␣r=1,2,3"
            raise Exception(et)

    def countCodesNumber(self):
        nc = Stripes.countCodes(self.k,
                                self.r)
        return nc
```

3) Matching codeword length to stripe width (height):
Having the stripe image resolution in the given direction `DD` (horizontal or vertical) and stripe size `L` in the same direction we get in function `matchCodeRank` minimum possible codeword length of the given repeatability *r*.

```
    @staticmethod
    def matchCodeRank(DD,L,r):
        for k in range(2,12):
            C = Stripes.countCodes(k,r)
            if L*C>=DD: return k
        return 0
```

E. Stripe image generator

Construction of stripe images is implemented by `generateStripeImages` static method.

Firstly, define its header.

1) start of method header:

```
    @staticmethod
    def generateStripeImages(
```

2) image horizontal resolution:

```
        W,
```

3) image vertical resolution:

```
        H,
```

4) stripe size:

```
        L,
```

5) code repetibility factor (with default):

```
        r=2,
```

6) orientation of stripes:

```
        orient='—',
```

7) if saving to file enabled (with default):

```
        to_file=True
```

8) end of method header:

```
    ):
```

Generating stripe images is based on code tables.

1) Firstly, prepare code tables:

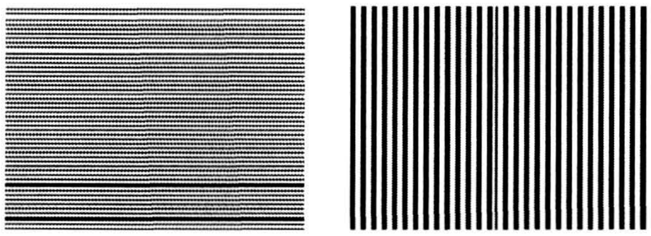
```
        D = H if orient=='—' else W
        k = Stripes.matchCodeRank(D,L,r)
        K = D/L
        self = Stripes(k,K,r=r)
        self.L = L
        self.simages = []
        prefix = "H" if orient=='—'\
                 else "V"
```

2) Next, for each image compose its stripes:

```
        for i in range(k):
            image = np.zeros((H,W),
                    dtype=np.uint8)
            o = 0
            for j in range(K):
                val = 255*self.codes[i][j]
                if orient=='—':
                  image[o:o+L,:] = val
                else:
                  image[:,o:o+L] = val
                o += L
            self.simages.append(image)
            if to_file:
                name = prefix+"_stripes_"
                name += str(i)+"_"
                name += str(self.k)+"_"
                name += str(self.r)
                matrixToImageFile(image,
                  "../image/"+name+".png")
        return self
```

3) We verify generated stripe images by watching them in the directory `../image/`:

```
sp_H = Stripes.generateStripeImages(
                    800,600,6,orient='—')
sp_V = Stripes.generateStripeImages(
                    800,600,7,orient='|')
```



F. Index image definition

Context of image index:

1) The 3D scene is illuminated in turn by *k* stripe images (patterns) of the same orientation.
2) Stripe camera images are registered in gray scale for each illumination.
3) Binarisation is performed according the formula (6)
4) For each pixel its codeword is converted to its index, i.e. the position of the codeword in the table of valid codes.

5) For visualization the index could be scaled to gray scale value or to an index of a pseudo colors table.

Building of index image:

1) Verify compatibility of the number of camera images to code rank:

```
def defineCodeIndexImage(self,cimages):
    k = self.k; r = self.r
    if len(cimages)!=k:
        et = ”Images␣incompatible”
        raise Exception(et)
```

2) Store all images into 3D table:

```
    H,W = cimages[0].shape
    table = np.zeros((H,W,k),
                dtype=np.uint16)
    for i in range(k):
        table[:,:,i] = cimages[i]
```

3) Perform in-place sorting with regard to 3-rd dimension:

```
    table.sort(axis=2)
```

4) Create the table of thresholds according one of four options:

```
    if self.th==1:
        thrs = (table[:,:,r-1]+\
                table[:,:,k-r])/2
    elif self.th==2:
        thrs = (table[:,:,0]+\
                table[:,:,k-1])/2
    elif self.th==3:
        thrs = np.zeros((H,W),\
                dtype=np.uint16)
        for i in range(r):
            thrs += table[:,:,i]
            thrs += table[:,:,k-i-1]
        thrs /= (2*r)
```

5) Build the code table:

```
    ctable = np.array(
                cimages[k-1]>thrs,
                dtype=np.uint16)
    for i in range(k-2,-1,-1):
        ctable *= 2
        ctable += (cimages[i]>thrs)
```

6) Create index table:

```
    index_table =\
            self.code_index[ctable]
    return index_table
```

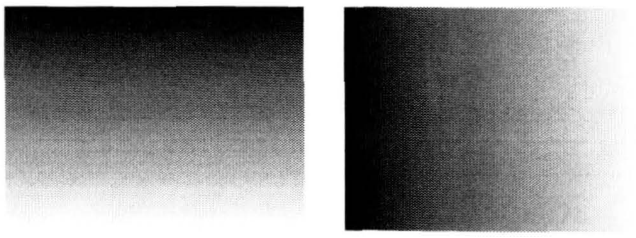
7) Testing index image generator.

   a) Define gray to rgb image convertor:

```
def toRGBImage(table):
    ntable = np.array(table,
                dtype=np.float32)
    mx = np.min(table)
    Mx = np.max(table)
    ntable -= mx*np.ones(
            table.shape,
            dtype=np.float32)
    ntable /= (Mx-mx)
    ntable *= 255.0
    gray = np.array(ntable,
            dtype=np.uint8)
    m,n=gray.shape
    rgb = np.zeros((m,n,3),
            dtype=np.uint8)
    for i in range(3):
        rgb[:,:,i] = gray
    return rgb
```

   b) For testing take stripe images with $k = 7, r = 2$, and check the result in the directory `../image`:

```
ci = Stripes.loadStripeFiles(7,2)
ind_H =\
    sp_H.defineCodeIndexImage(ci)
rgb_image = toRGBImage(ind_H)
matrixToImageFile(rgb_image,
”../image/index_test_image_H.png”)
```



G. Importing stripe images from files

```
@staticmethod
def loadStripeFiles(k,r,orient='—'):
    cimages = []
    prefix = ”H” if orient=='—'\
            else ”V”
    for i in range(k):
        name = prefix+”_stripes_”
        name += str(i)+”_”
        name += str(k)+”_”
        name += str(r)+”.png”
        image = imageFileToMatrix(
                '../image/'+name)
        cimages.append(image)
    return cimages
```

H. Virtual corners extractor

An overlay of vertical and horizontal index images registered by a stripe and video camera produces a virtual quadrilateral mesh which can be used either for the system calibration with regard to to depth information or during on-line 3D free-point views creation.

Therefore stripe detection in camera image for illuminated scene is of crucial importance. For calibration intersections of stripe edges, so called corners are also useful.

For corners and stripe edge detection the popular method is Harris filter which is a detector of abrupt changes in a natural image. However, the index image is a map of code image. The nature of stripe code is its stability within the stripe. The changes within the stripe mean error code detection inside the stripe. Moreover, changes between stripes in calibration planar scene are equal to one, and small integer for natural scenes. Therefore another approach based on computing scaled variance in 3×3 window of index image seems more adequate, as the variance within the stripe is equal to zero:

$$\mathrm{svar}_I(p) \doteq 8 \sum_{0<\|q-p\|_1<3} I(q)^2 - \left(\sum_{0<\|q-p\|_1<3} I(q) \right)^2$$

The following properties can be proved: *If the pixel p belongs to a stripe edge, its 3×3 neighborhood there are $k \in [1,7]$ pixels indexed to i, and $8-k$ pixels has got the index $i+1$ then the scaled variance depends only on k accepting only four different values* $56, 96, 120, 128$:

$$svar_I(p) = 8k(8-k), \ k = 1, \ldots, 7$$

Relevant summation will be performed by the function `f2d`, an alias of the function `__filter2d__` from the module `mfilter`. There are two kernels (masks) in the implementation:

$$\texttt{mask} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \ \texttt{mask2} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

1) Parameters:
   a) start of the header:

```
@staticmethod
def extractCorners(
```

   b) index image for horizontal stripes:

```
    ind_h,
```

   c) index image for vertical stripes:

```
    ind_v
```

   d) end of the header:

```
):
```

2) Compute sums of indexes and their squares for each pixel of index images:

```
    mask = [[1,1,1],[1,0,1],[1,1,1]]
    sum_h = f2d(ind_h,kernel=mask,
                dtype=np.int32)
    sum_v = f2d(ind_v,kernel=mask,
                dtype=np.int32)
    sum2_h = f2d(ind_h*ind_h,
                     kernel=mask,
                 dtype=np.int32)
    sum2_v = f2d(ind_v*ind_v,
                     kernel=mask,
                 dtype=np.int32)
```

3) Detect pixels with positive scaled variance in both index images:

```
    J_0 = np.array(
          ((sum2_h<<3)>sum_h*sum_h)\
          &((sum2_v<<3)>sum_v*sum_v),
          dtype=np.uint8)
```

4) Leave only groups of pixels having at least three elements:

```
    J_1 = f2d(J_0,kernel=mask,
              dtype=np.uint8)
    J = np.array(J_1>2,
                dtype=np.uint8)
    y,x = np.where(J)
```

5) Sum up indexes around detected elements:

```
    mask2 = [[1,1,1],[1,1,1],[1,1,1]]
    sum_h = f2d(ind_h*J,kernel=mask2,
                dtype=np.int32)
    sum_v = f2d(ind_v*J,kernel=mask2,
                dtype=np.int32)
    sh = sum_h[y,x]; sv = sum_v[y,x]
```

6) Determine group centroids:

```
    yres,xres = ind_h.shape
    X = np.outer(
        np.ones(yres,dtype=np.int32),
        np.arange(xres,dtype=np.int32))
    Y = np.outer(
        np.arange(yres,dtype=np.int32),
        np.ones(xres,dtype=np.int32))
    sum_X = f2d(X*J,kernel=mask2,
                dtype=np.int32)
    sum_Y = f2d(Y*J,kernel=mask2,
                dtype=np.int32)
    sx = sum_X[y,x]; sy = sum_Y[y,x]
```

7) Correct the number of pixels in the neighborhood and return mean values:

```
    nxy = J_1[y,x]+np.ones(len(y),
            dtype=np.float32)
    return sy/nxy,sx/nxy,sh/nxy,sv/nxy
```

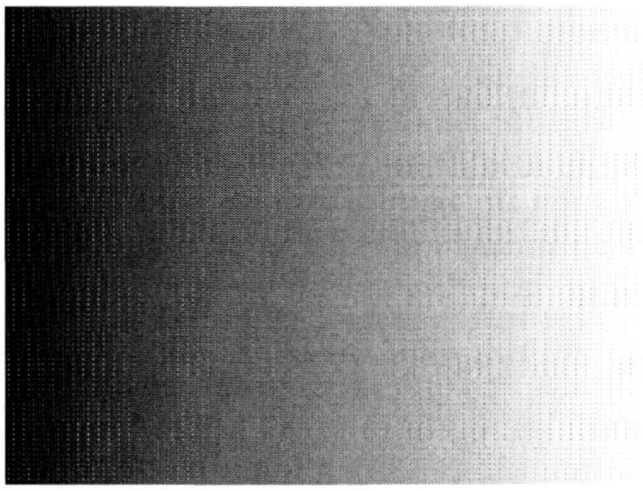
I. Testing corner extractor

To get vertical index image we define eight stripe images with $r = 2$:

```
ci = Stripes.loadStripeFiles(8,2,
                             orient='|')
ind_V = sp_V.defineCodeIndexImage(ci)
rgb_image = toRGBImage(ind_V)
matrixToImageFile(rgb_image,
"../image/index_test_image_V.png")
y,x,h,v = Stripes.extractCorners(ind_H,ind_V)
```

To watch the results, the detected corners are indicated by red color in vertical index image:

```
y = np.array(y+0.5,dtype=np.int32)
x = np.array(x+0.5,dtype=np.int32)
print(y[:500],x[:500])
rgb_image[y,x,0] = 255
rgb_image[y,x,1] = 0
rgb_image[y,x,2] = 0
matrixToImageFile(rgb_image,
"../image/extractor_test_image.png")
```



In the figure perfect corner extraction is observed. This is possible only if perfect image binarization takes place what in real camera images is not true. Copying with imperfection is the main research goal. In this case, we simply avoid detection of stripe corners in favour of stripe inner pixels which exhibit much higher success rate of detection.

V. Conclusions

The conclusions can be I divided into positive and negative parts. On positive side we have:

1) Clear representation of top-down software design.
2) Full integration of design with implementation.
3) Nonlinear track of code development.
4) Easy integration of main path of code design with testing path.
5) Tutorial character of system manual which includes algorithms, formulas, resulting images, and the system code, too.

On the negative part we get invisible (in the paper) drawbacks:

1) In Latex editor there is no support for syntax check of programming languages. Therefore, the person who is both the programmer and the designer has to have perfect knowledge of the programming language and the libraries used.
2) The unknown algorithms, just developed or modified are hard to write as the cycle of editing (in a Latex editor), integration (by the Integra application), and testing (in a development environment, e.g. IDLE) gives the significant time overhead.

References

[1] P.Fechteler and P. Eisert, Adaptive Colour Classification for Structured Light Systems, IET Journal on Computer Vision, vol. 3, no. 2, (June 2009) 49-59
[2] S. Zhang and P. S. Huang, High-resolution Real-time 3-D Shape Measurement. Optical Engineering, 45(12), (2006).
[3] J. Pages, J. Salvi, C. Collewet, and J. Forest, Optimised De Bruijn patterns for one-shot shape acquisition, Image and Vision Computing, 23:707720 (2005).
[4] M. Pharr, G. Humphreys, Physically Based Rendering, Morgan-Kauffman, 2004.
[5] J. Salvi, J. Pages, and J. Batlle, Pattern codification strategies in structured light systems, Pattern Recognition, 37:827849, 2004.
[6] J. Ghring, Dense 3-d surface acquisition by structured light using off-the-shelf components, Videometrics and Optical Methods for 3D Shape Measurement 4309 (2001) 220231.
[7] E. Horn, N. Kiryati, Toward optimal structured light patterns, Image and Vision Computing 17 (2) (1999) 8797.
[8] E. Trucco, R. B. Fisher, A.W. Fitzgibbon, D. K. Naidu, Calibration, data consistency and model acquisition with laser stripers, International Journal Computer Integrated Manufacturing 11 (4) (1998) 293310.
[9] R. J. Valkenburg, A. M. McIvor, Accurate 3d measurement using a structured light system, Image and Vision Computing 16 (2) (1998) 99110.
[10] D. Caspi, N. Kiryati, J. Shamir, Range imaging with adaptive color structured light, Pattern analysis and machine intelligence 20 (5) (May 1998) 470480.
[11] O. Faugeras, Three-Dimensional Computer Vision," MIT Press, 1993.
[12] D. Knuth, The TeXbook, Addison Wesley, 1984.
[13] J. L. Posdamer and M. D. Altschuler, Surface measurement by space-encoded projected beam systems, Comput. Graph. Image Processing 18(1), 117 (1982).
[14] http://www.literateprogramming.com/
[15] http://en.wikipedia.org/wiki/Structured_light

SESSION 1:
IMAGE PROCESSING I

Level-Set Image Processing Methods in Medical Image Segmentation

Marcin Maciejewski MSc
Wojciech Surtel PhD.
Institute of Electronics
Lublin University of Technology
Nadbystrzycka 38a 20-618 Lublin
m.maciejewski@pollub.pl

Teresa Małecka-Massalska M.D PhD.
Department of Human Physiology
Medical University of Lublin
Radziwiłłowska 11, 20-080 Lublin

Abstract — **In this paper two image processing methods for use in medical image processing based on the level set method are described. The theoretical basics are described and the methods are applied to a set of sample CT images. The results are then compared.**

Keywords — image processing, level set, medical diagnosis

I. Introduction

Image processing and recognition is an area of interest of many scientists and has many applications. One of these is a vital part of diagnosis of different kinds of cancer. It is a disease that manifests as a palpable growth in the body and as such it can be discovered using processing of images from many techniques. Examples are CT scans, MRI images, X-rays and skin lesion photography[10]. Those images carry a large amount of valid information, making them a vital base for diagnoses. Unfortunately, in most cases the diagnostic process requires a skilled human operator to assess the data. Having in mind that the images are often distorted by noise and that the diagnosis itself is seldom based on fixed rules it is obvious, that the process can sometimes give inconclusive results. Another problem is that in many cases it is imperative to administer the patient with a contrast solution to acquire good quality images, which in some patients can cause negative effects depending on their physiology. Successful treatment by removing the growth cannot be performed without first knowing the exact area to operate in. It makes the surgeon life easier and also decreases the chance of leaving some diseased cells that can spread and regrow. The fact that it is one of the major causes of death around the world[9] makes the research in the field very important. Also, first symptoms of the disease can be hard to connect to their actual cause, which makes the development of tools for diagnostic aid a great benefit to medicine.

II. Level set methods

In the study we compared two image segmentation methods. One of these was the Distance-Regulated Level-Set method, the other one was the Chan-Vese method with and without implementing optimization for vector images. Both methods are using the level set method as a base to describe the transformation of the contour.

The level-set method is a well known and appreciated technique for numerical description of shapes in the Cartesian system. It's advantage over other methods is that it allows for faster computation without parametrization of shapes, which in turn improves efficiency. This makes the method suitable for describing dynamic changes in objects, like contour evolution in image segmentation. In sequential iterations new contours are being calculated, resulting in a level set[1]. The method was first proposed and introduced by Osher and Sethian[2][3]. In the method we describe a certain area Ω with an edge Γ. The velocity of the edge v between steps depends on the position, shape, time and external conditions. It is possible to conduct calculation in the x domain after discretization. We are trying to define a function ϕ(x,t), where x is the position in Cartesian space and t is time, describing the moving contour (fig.1). The level set method can be used in different applications[4]. In general, the algorithm can be described using a block diagram like the one in figure 2. The contour stops moving when a certain condition of minimal energy is met. It is possible to choose the external energy functional to best fit the task. The distance regulated level set method utilizes a functional (1) introduced in [4]:

$$E(\phi)=\mu * R_p(\phi)+ \lambda * L_g(\phi)+ \alpha * A_g(\phi) \qquad (1)$$

where:
E(ϕ) is the energy,
Rp(ϕ) is the level set regularization term,
Lg(ϕ) is minimized when the zero level contour of is located at the object boundaries,
Ag(ϕ) is introduced to speed up the motion of the zero level contour in the level set evolution process.

The Chan-Vese method differs in the sense that it implements another approach to defining and minimizing the energy functional(2).

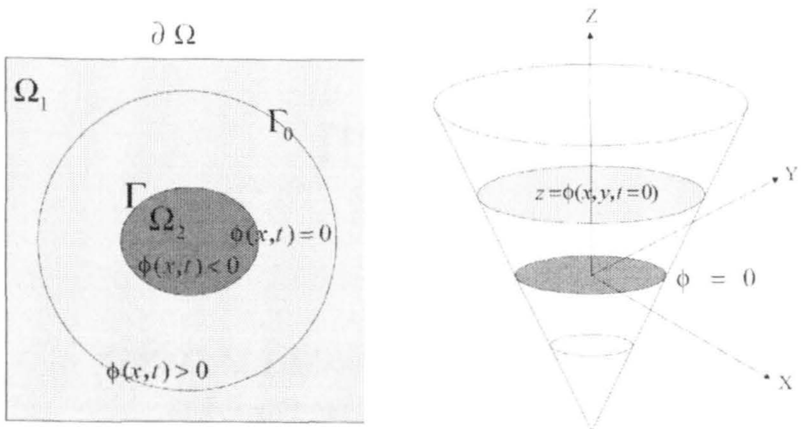


Figure 1. The level set function [1].

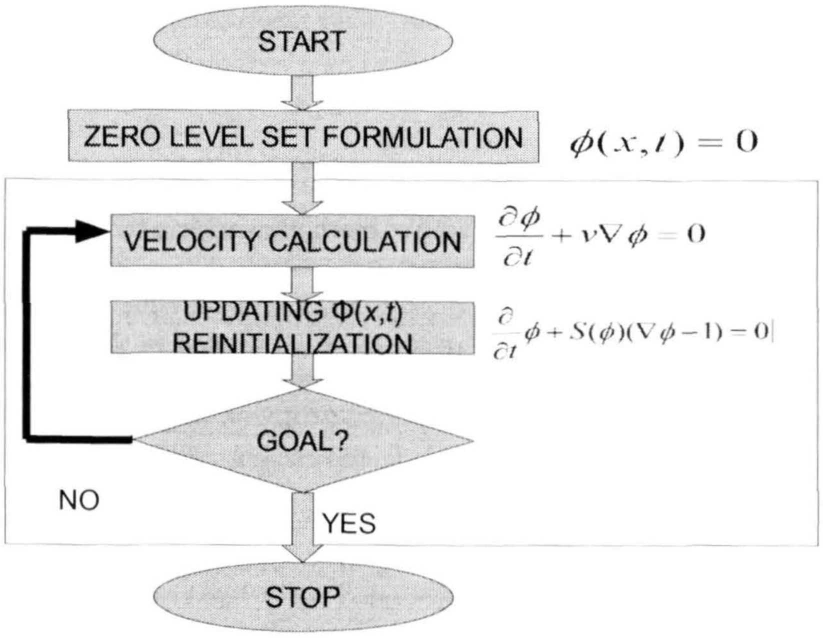

START
ZERO LEVEL SET FORMULATION
$\phi(x,t)=0$
VELOCITY CALCULATION
$\frac{\partial\phi}{\partial t}+v\nabla\phi=0$
UPDATING Φ(x,t)
REINITIALIZATION
$\frac{\partial}{\partial t}\phi+S(\phi)(\nabla\phi-1)=0$
GOAL?
NO
YES
STOP

Figure 2. General image segmentation algorithm using the level set function.

$$
\begin{aligned}
E(\phi)=&\mu\Big(\int_{\Omega}\|(\nabla H(\phi))\|dx\Big)^{p}+\gamma*\int_{\Omega}\nabla H(\phi)dx\\
&+\lambda_1*\int_{\Omega}(\|(I-c_1)\|)^2\nabla H(\phi)dx+\lambda_2*\int_{\Omega}(\|(I-c_2)\|)^2\nabla(1-H(\phi))dx
\end{aligned}
\tag{2}
$$

Where:

E(F) is the energy,

First segment is the weighted term from the total contour length, it's impact can be increased for smoother contour,

Second segment defines the total area inside the contour,

Third and the fourth segment are proportional to the variance of the pixel variance inside and outside the contour respectively.

In figure 3 an example segmentation result is shown.

During the segmentation, the acceleration and velocity of the contour are varying depending on the moment. While the contour φ(x,t) is far from the solution, the velocity and the acceleration increases. When the contour approaches the desired area, the acceleration and speed decrease. Acceleration and velocity chart in an example segmentation is presented in figure 4.

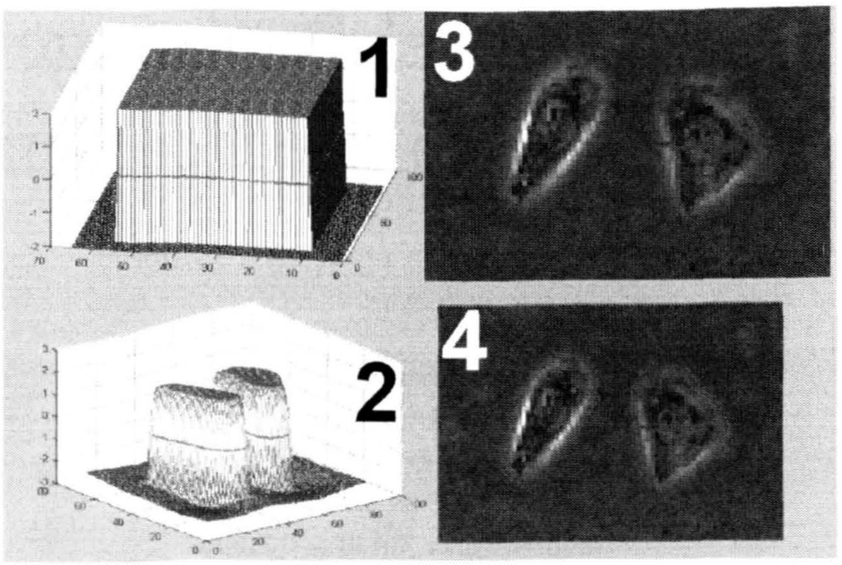


Figure 3. Example segmentation of an image of two cells[4]. The original level set 1 and image 3 are transformed to the desired level set 2 and gives contour function in the image 4.

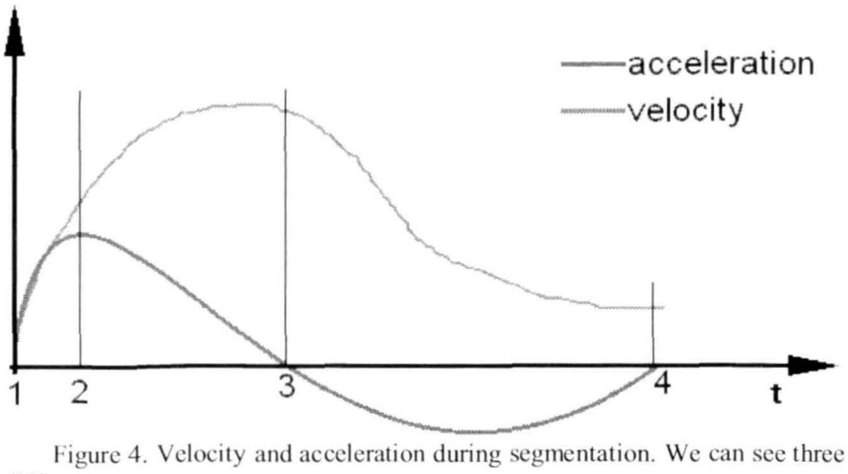


Figure 4. Velocity and acceleration during segmentation. We can see three different areas

1 to 2 – far from the contour,

2 to 3 – near the contour,

3 to 4 – slowing down in the intermediate vicinity of the contour, approaching end condition with zero velocity.

III. DATA

In the course of research, we processed a set of medical images using the two methods. The images were obtained from the Second Radiology Department in Lublin. The images were taken using CT scans. We have chosen 16 images representing the scans of the patients' torso and two side views. The images were all gray scale and their size was 600x600 pixels. On the images, a few distortions in the form of markers and descriptions were present. The images were delivered in .JPG format. The authors of the original algorithms in MATLAB were Chunming Li[4] and Yue Wu. In the course of research, the codes were modified to allow for batch image processing, easy result interpretation, evolution speed and acceleration calculations. We applied the methods to both noisy and Gaussian filtered images, and both with and without the markers. In some cases, artificial noise was added to test the methods' robustness. Number of iterations was recorded for all the cases and overall quality of segmentation versus time taken was assessed. The accuracy of the segmentation was assessed by a MRI diagnostician.

IV. Results

The consequent curves are usually chosen from the set calculated using the method of curve diffusion[4]. Proper definition of end conditions has a large impact on the segmentation results. When too many iterations are performed, some details of the image can be treated as noise and removed. In the same way, after too little iterations the resulting mask can be nonspecific. It is shown in figure 5. Usually, the number of iterations doesn't impact the center of mass position of the objects. A major difference can be observed in the object's longest diameter, it can decrease as soon as after 70 iterations. This often results in another distance being chosen as the longest diameter. It is shown in figure 6. The longest diameters were calculated between two furthest pixels. The surface areas were calculated by counting the number of pixels in each object. The Distance-Regulated method only gave good results for noiseless images of low complexity. In some cases, it was impossible to achieve good results despite using noise filters and many iterations. The method used more resources for each calculation and was not preferable. In figure 7 example segmentations of a noisy image using both methods are shown. Due to the fact that the methods differ in calculating the ending condition, the Chan-Vese method performed much faster.

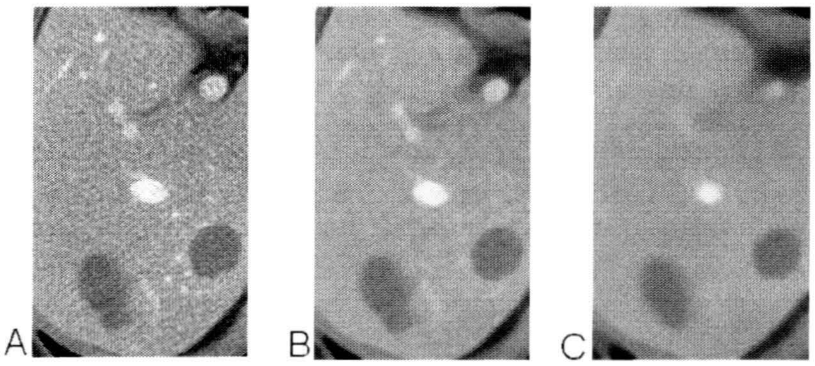


Figure 5. Tomographic image of the liver. a) original image, b) image processed by 50 iterations of curve diffusion, c) image processed by 150 iterations of curve diffusion.

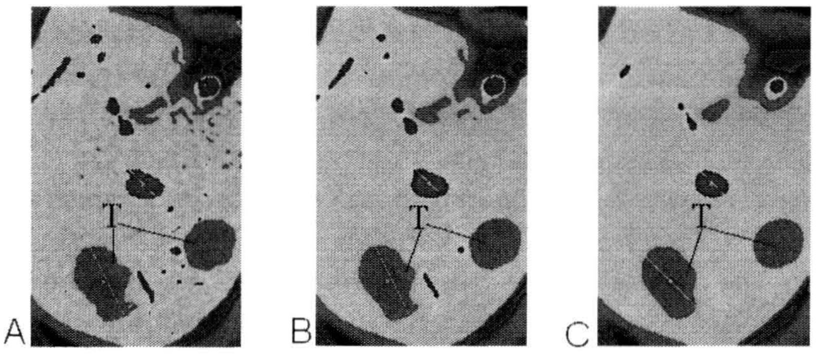


Figure 6. An example transformed liver image with object center mass and longest diameter shown. Segmentation with curve diffusion using a)20, b)50 and c) 150 iterations. The blue areas represent blood vessels. The bottom two red areas marked with "T" are tumor masses.

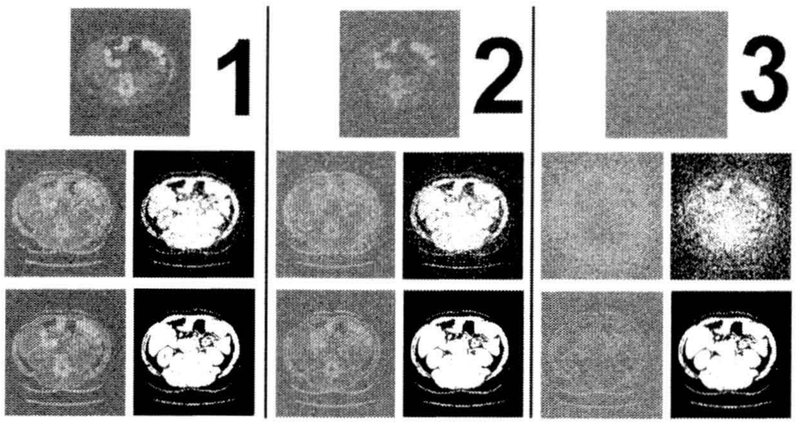


Figure 7. Segmentation of a noisy image with salt-and-pepper noise. In the top row, images with noise. In the middle row, the Distance-Regulated method, in the bottom row the Chan-Vese method for vector images. Segments 1, 2 and 3 correspond to 0.2, 0.5 and 1.0 SNR accordingly.

V. Conclusion

The Distance-Regulated method returns satisfactory results only for simple and clean images and requires much more resources. During the segmentation process the Chan-Vese method performed faster and gave satisfactory results. All the masks were accurate and the number of iterations never exceeded 300. The method for vector images proved to be insensitive to noise and performed well for Gaussian and salt-and-pepper noise. The method can successfully be used to obtain segmentation masks even from noisy images and offers fast computation times. It has potential in the field of medical imaging and can be used to show significant masses.

References

[1] Tomasz Rymarczyk, "Zastosowanie metody zbiorow poziomicowych w tomografii impedancyjnej", doctor's thesis, Warsaw 2010J.

[2] Osher S., Fedkiw R.: "Level Set Methods and Dynamic Implicit Surfaces". Springer, New York 2003.

[3] Sethian J.A.: "Level Set Methods and Fast Marching Methods". Cambridge University Press 1999.

[4] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D. Fox "Distance Regularized Level Set Evolution and Its Application to Image Segmentation", IEEE Transactions on Image Processing, vol. 19, no. 12, December 2010

[5] Pham, Dzung L.; Xu, Chenyang; Prince, Jerry L. (2000). "Current Methods in Medical Image Segmentation". Annual Review of Biomedical Engineering 2: 315-337

[6] L.S.S.Reddy, Ramaswamy Reddy, CH.Madhu & C. Nagaraju, "A novel image segmentation technique for detection of breast cancer" International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 201-204

[7] "Tenn Francis Chen Medical Image Segmentation using Level Sets" Technical Report #CS-2008-12

[8] M. Droskey, B. Meyerz, M. Rumpfy, K. Schallerz, "An adaptive level set method for medical image segmentation", Institut für Angewandte Mathematik, Klinik für Neurochirurgie, Universität Bonn

[9] American Cancer Society. "Cancer Facts & Figures 2010". Atlanta: American Cancer Society; 2010.

[10] Dimitrios H.Roukos M.D., and Niki J.Agnantis M.D. "Gastric Cancer: Diagnosis, Staging, Prognosis", Gastric Breast Cancer 2002; 1(1): 7-10

Computer Simulation of Magnetic Resonance Angiography Imaging. Parallel Implementation in Heterogeneous Computing Environment

Artur Klepaczko, Piotr Szczypiński, Grzegorz Dwojakowski, Marek Kociński, Michał Strzelecki
Lodz University of Technology
Institute of Electronics
90-924 Lodz, ul. Wolczanska 211/215
Email: aklepaczko@p.lodz.pl

Abstract — **Magnetic resonance imaging (MRI) is currently widely used in medical image diagnosis. However, MR scanners are extensively used in clinics and thus are rarely accessible for experimentation. In consequence, the number of images available for image processing methods evaluation is too low and there appears a need for a method to generate synthetic images. In their previous works, the authors studied various methods for blood vessels segmentation and tracking. Effectiveness of the designed algorithms requires objective verification which implies repetition of experiments for large number of images and comparing the results with some ground truth models. Therefore, this study aims at designing a computer system which implements numerical routines for generation of synthetic MRA images. In particular, in this paper we study the performance of various configurations of assembled computer grid and analyze their potential in angiographic image synthesis.**

I. Introduction

Magnetic resonance imaging (MRI) is currently widely used in medical image diagnosis. This non-invasive technique allows acquisition of high-resolution volumetric images which visualize even small details of biological tissues precisely positioned relative to the interior of the whole body. On the other hand, recent advances in the development of image processing algorithms facilitates their usage in quantitative anlysis of MR images. If conducted properly, such analysis can augment objectivity and correctness of the medical diagnosis. However statistically credible validation of the designed methods is hampered by the significant cost of collecting MR images solely for the research purposes. Additionally, MR scanners are extensively used in clinics and thus are rarely accessible for experimentation. In consequence, the number of images available for evaluation studies is too low and there appears a need for a method to generate synthetic images. One of the approaches that helps to overcome these impediments is computer simulation of magnetic resonance imaging [1].

In particular, we put focus on magnetic resonance angiography (MRA) directed at visualization of the human vessel system. In their previous works [2], [3], the authors studied various methods for blood vessels segmentation and tracking in 3D MR images acquired using Time-Of-Flight (ToF) or Susceptibility Weighted Imaging (SWI) sequences. Effectiveness of the designed algorithms requires objective verification which implies repetition of experiments for large number of images and comparing the results with some ground truth models.

Therefore, this study aims at designing a computer system which implements numerical routines for generation of synthetic MRA images. This general aim can be decomposed into several subtasks which include:

1) definition of the virtual object containing synthetic vessel branches and characterized not only by the magentic properties (such as spin-spin and spin-lattice relaxation time constants, hydrogen proton density, tissue magnetic susceptibility) but also by the hydrodynamic parameters of a substance flowing through virtual vessels;
2) blood flow simulation which generates information about displacement of moving voxels during MR image formation;
3) simulation of electromagnetic phenomena which take place in real MR scanner spin selective excitation, precession, relaxation, signal induction, etc.
4) image formation k-space filling, filtering in the spatial-frequency domain, application of the Fourier transform.

The above described tasks are computationally extensive, especially if the target image is a 3D high-resolution data structure. Most existing MRI simulators are implemented using various architecture computer grids. Thus, for the need of our project and as a first step towords the target system, we designed our own parallel implemention of the core MRI simulator, i.e. no flow effects are taken into account. The objective was to use the available resources and thus gain guaranteed, flexible, on-demand access to it in any future exeperiments and be independent from publicly available, but of limited access, open grid projects. As a consequence, our runtime environment is developed as a heterogeneous grid of UNIX/LINUX machines which communicate with one another over local ethernet network using the SSH protocol.

In this paper, we study the performance of various configurations of our computer grid and analyze its potential in angiographic image synthesis. In the reminder of this paper

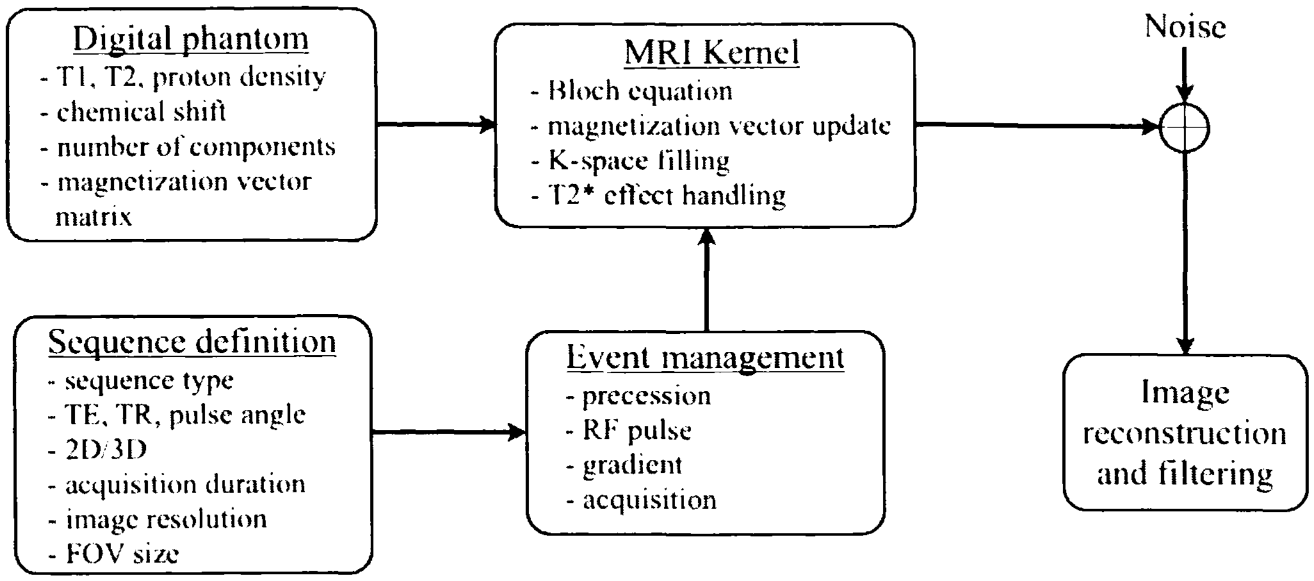


Fig. 1. Main software modules of the MRI simulator

we describe the overall concept of the MRA simulator and main parts of the system (Sect. II). Then in Sect. III, we provide here the description of how we plan to deal with the flow phenomena, so that angiographic images can be simulated. In Sect. IV we give technical specification of the computer grid used in the experiments, as well as the details of simulator parallel implementation based on the MPI standard. The performance tests are reported in Sect. V and finally Sect. VI concludes.

II. MRI Simulation

Difficulties in acquistion of large volumes of real MRI images solely for research purposes constitute the main reason why over the last two decades there appeared a number of attempts to build computer MRI simulators. The proposed solutions [4], [5], [6], [7] differ from one another by the approach to modelling tissue of imaged objects, the MR image synthesis routines, and the degree to which different artifactual or undesired phenomena (like noise, chemical shift, ringing or Gibbs phenomenon, partial volume effect etc.) are taken into account.

In the approach proposed in [8], the imaged objects are defined as parameter maps. These maps visualize distribution of T1 (spin-lattice relaxation time constant), T2 (spin-spin relaxation) and rho (proton density) values within the object. These values are determined based on real images. Formation of new ones is performed using different repetition and echo times.

On the other hand, in paper [9], the virtual organ object is modelled using the concept of spin density. By the use of inverse Fourier Transform, the simulator constructs the image representation in the spatial-frequency domain (k-space formalism). Simulation of the signal sampling phase and specific imaging sequence is accomplished through appropriate selection of the k-space elements and their amplitude and phase alteration. It is raised that this approach requires separate simulation phases for each of the modelled tissues. This introduces significant impediment if a voxel contains more than one tissue (i.e. partial volume effect).

In comparison with other approaches, more realistic simulation effects are achieved by solutions proposed in [5], [6], [7]. Models of the imaged objects used therein require for each object voxel and every tissue component definition of T1, T2, and rho parameters, as well as value of the frequency offset linked to chemical shift artefact. In order to deal with the partial volume effect it is sufficient to determine how much of particular tissue material is present in a voxel.

The simulator used in our study is based on the SIMRI (fr. *SIMulateur de MRI*) project described in [5], due its versatality and proven effectiveness in producing realistic MR images. Moreover, it seems to be best suited – after necessary modifications – to our proposed angiography simulator. The main software modules of the originally proposed system – reimplemented by our team – are presented in Fig. 1.

A. Digital phantom description

The input to the simulator is the virtual object composed of 3D rasterized components, each of which corresponds to a different tissue type. Tissues are described by their spin-lattice (T1) and spin-spin (T2) relaxation times, as well as proton density (ρ) values. Every component is also described by the water-relative frequency shift which allows simulation of the chemical shift artifact. This value is determined assuming that the main magnetic field B_0 = 1T. Moreover, each object voxel is associated a 3D magneitzation vector, whose state at subsequent stages of image formation is controlled by the MRI simulation kernel. In addition to raster size (number of voxels in x, y and z coordinates), the object is characterized by its physical dimensions. Thus, every object voxel accomodates information about portion of the tissue component it contains, (T1,T2,ρ)-tuples for each component, and current magnetization vector state being a sum of magnetization vectors calculated for every tissue.

B. Sequence definition

The image synthesis procedure is controlled by the group of paramaters which set up sequence type and output image contents. The latter refers both to the image resolution which can be lower or equal to the input virtuual object raster size and to the position and size of the field of view (FOV). By defualt FOV embraces whole object, but it can be reduced to only a part of it and its center can be moved to any object point. This is important to keep the record of the FOV center offset and its location relative to gradients isocenter, since it affects the gradient value experienced by each voxel.

The sequence definition is mainly determined by the sequence type, which cen be either spin echo (SE) or gradient echo (GE) imaging. In both cases, the program requires specification of the echo and repetition times (TE and TR accordingly). In the case of GE, it is also necessary to define the flip angle (α). The sequence type also include information whether the target image is 2- or 3-dimensional. For the 2D image, one has to specify the slice position (x coordinate). Eventually, the signal sampling window is determined by the acquisition time which must ensure that the sampling frequency meets the Nyquist theorem.

C. Event management

The event management module is responsible for invoking subsequent steps of computations based on the current simulation stage. It begins with establishing the initial magnetization of the object by allowing all of its spins to freely precess in the presence of the main external magnetic field B$_0$. Then, depending on the sequence type, the object magnetization state is altered through the application of an RF pulse, free precession, refocussing pulse (SE imaging only) and application of phase encoding gradients (in y and z directions in the case of volumetric imaging or only in y direction for the 2D case). Eventually, the frequency encoding gradient is applied and the signal acquisition step is executed. At each step the MRI kernel is invoked with the appropriate parameters.

D. MRI kernel

The heart of the whole system is the MRI kernel which implements the discrete time solution to the Bloch equation [10]. It uses the rotation matrices and exponential scaling to modify the voxel magnetization vectors in accordance to the specified sequence events. For the details of kernel implementation we refer the reader to the original paper [5] – here we recall only the main bits of this system module.

First of all, the magnetization vector $\vec{M}$ at location $\vec{r} = [x, y, z]$ is given by

$$\vec{M}(\vec{r}, t + \Delta t) = Rot_z(\theta_g) Rot_z(\theta_i) R_{\text{relax}} R_{\text{RF}} \vec{M}(\vec{r}, t), \quad (1)$$

where $Rot_z(\theta)$ rotates the magnetization vector around the z-axis in reply to the phase encoding gradient (θ_g) and as a consequence of the main magnetic field inhomogeneity (θ_i). The R_{relax} rotation matrix is responsible for the relaxation effects and is given as

$$R_{\text{relax}} = \begin{bmatrix} e^{-\frac{\Delta t}{T_2(\vec{r})}} & 0 & 0 \\ 0 & e^{-\frac{\Delta t}{T_2(\vec{r})}} & 0 \\ 0 & 0 & 1 - e^{-\frac{\Delta t}{T_1(\vec{r})}} \end{bmatrix}. \quad (2)$$

Eventually, the R_{RF} encapsulates the effect of the RF excitation pulse which flips the magnetization vector around an axis belonging to the xy-plane by the angle in time Δt. In the presence of gradient, the effective flip angle is α' and R_{RF} is calculated as

$$R_{\text{RF}} = Rot_z(\phi) Rot_y(\beta) Rot_x(\alpha') Rot_y(-\beta) Rot_z(-\phi), \quad (3)$$

where

$$\alpha' = -\Delta t \sqrt{(\Delta\omega)^2 + \left(\frac{\alpha}{\tau}\right)^2},$$

$$\beta = \tan^{-1}\left(\frac{\Delta\omega}{\alpha/\Delta t}\right),$$

and $\Delta\omega$ is the local frequency offset from the main static magnetic field induced by gradients, RF pulse and field inhomogeneities.

Signal acquistion is performed by aggregating the transverse magnetization components over all object voxels and this summation is repeated the number of times needed to fill one K-space line. Between each two acquistions, the voxel magnetization states are updated due to relaxation effects that take place during the sampling period δt. Thus, one point $s(t)$ of the K-space line is calculated as follows

$$s(t) = \sum_{\vec{r}} \vec{M}(\vec{r}, t)\vec{x} + j \sum_{\vec{r}} \vec{M}(\vec{r}, t)\vec{y}. \quad (4)$$

The MRI kernel is actually much more versatile than described here. It offers T2* effect handling, variety of pulse wave forms, gradient crushing, numerical transverse magnetization spoiling, can take into account various artefatcs linked to static field inhomogeneities, field default intensity values or tissue susceptibility. Apart from standard SE and GE sequences, there are more sophisticated sequences implemented, such as e.g. True-FISP technique (Fast Imaging with Steady-state Precession). It also allows adding noise to the K-space data, although in the current project phase, this feature is restricted to white Gaussian noise only, which is a simplification of the true MR noise characteristics. As indicated in [11], the better model here would incorporate the Rice function.

E. Image reconstruction and filtering

The last stage of image formation procedure consists in transformation of the raw data in the spatial-frequency domain (K-space) into image intensity domain. This is accomplished by the fast Fourier transform, followed by calculation of the magnitude of – in general – complex tranformation output. Before application of the FFT routine, it may be necessary to filter the raw data to reduce the Gibbs artifact in the case of small resolution images. We decided to port the filtering routine (as welll as FFT-based image reconstruction) to the Matlab environment and thus we can use any kind of low-pass digital filter available in the Signal Processing Toolbox [12].

III. Extension to angiography imaging

The major modification we need to introduce to the SIMRI simulator concerns the virtual object, which now must include information about replacement of the flowing blood. We decided to determine this information using a separate program dedicated to flow simulation. For that purpose we employed the COMSOL Multiphysics 4.2a software [13]. In our initial experiments we used a simple digital flow phantom, composed of a single cylindrical tube of impermeable boundaries placed in porous material, able to absorb arbitrary volume of liquid. Since under the framework of this paper we are mostly interested in MR image formation process, the flow phenomena are not covered here and will be discussed in a separate work.

It must be noted, that the flowing blood particles move at different speed, depending on their position within a vessel and the vessel size. Replacement of the slower particles during a single simulation time step (such as a sampling period e.g.) may be less than a voxel. This breeds a need for even a deeper remodelling of the digital phantom than simple addition of per-voxel flow information. The two feasible solutions for this issue is to perform computations at sub-voxel level or redefinition of the object (at least for the moving component) so that instead of voxels arranged in a regular lattice it is composed of particles, each storing its current position within the object volume. Because the prior startegy does not solve the problem entirely but only reduces it to the limit of the voxel quantization scale, we decided to focus on the second approach.

The model of object constructed as a set of particles requires that each particle is associated a portion of media – call it a cell – it represents. Contribution to the output image coming from each separate cell should be proportional to its size. Therefore, after each time step and magnetization vectors update, our system is designed to recalculate new mesh of simplices (cells) linked to the moving particles. We chose the 3D Delaunay triangulation [14] method to perform this operation. The triangulation is invoked on the points corresponding to current particle locations and static points placed all over the vessel walls.

The signal acquistion procedure for the moving liquid compnent is executed similarily to stationary tissues, but now the sum in Eq. (4) is taken over cells and not over regularly distributed voxels. Moreover, each element of the sum is weighted by the factor which relates volume of a cell to the volume of an output image voxel. Note, that the K-space structure remains a regular grid of voxels, as the signal sampling procedure is independent from the object internal definition. Eventually, the collected K-space data are simply added to complex signal volumes built for the stationary tissues.

Extension to angiography imaging sequences is still under development. Thus, the rest of the paper presents the performance of the designed cluster environment emplyed for the synthesis of basic MRI images.

IV. Cluster environment

As noted in the introduction, we aimed at constructing our own computer cluster built on available hardware resources instead of utilizing any of the exisiting and publicly avaialble remote grid projects. In the latter case, the access to grid execution units is subject to specific rules which define job submission scheduling process, software interface for parallel programming, total amount of working hours granted to a user, etc. In the current phase of our project, which is still under intensive development, it is preferable to have on-demand access to effiicient computing enviroment allowing for instant testing of any of the incremental and major system modifications.

A. Cluster configuration

The network of computers used in our simulations is composed of UNIX and LINUX machines of various architectures. The grid is composed of two server units (both possesing a pair of Intel quad-core Xeon CPUs), one 20 iMac (with double-core Intel Core Duo), one 27 iMac (quad-core Intel Core i7) and 6 MacMini computers, each equipped with an Intel Core i5 processor. Enabled with the hyperthreading feature, the i7 CPU is actually able to run 8 threads in parallel, while double-core i5 can run 4 threads simultaneously. This gives effectively 50 cores of 134,4 GHz total computing power and potentially 537.6 GFlops of peak performance. Note, that both server computers belong to the i386 architecture, while other units to the x86_64 platform. This variation in operating systems and hardware configuration makes the designed cluster a heterogeneous computing environment. As a consequence, flexibile communication protocol, transparent for both Linux and Mac OS X machines, is needed to ensure efficient and reliable parallel job execution. In conjuction with the Open MPI library that we used to parallelize the simulator, it occurred that the SSH protocol provides the easiest way to establish the grid.

B. Parallelization using Open MPI library

The Message Passing Interface (MPI) is the well-recognized standard in the domain of high performance computing. It is specifically designed for use in cluster/grid systems with distributed memory model, i.e. where each processing unit has its own memory addressing space. The MRI simulator developed under the SIMRI project was based on the MPICH-G2 [15] distribution. However, we decided to build our implementation around the Open MPI library [16] due to its better integration with Mac OS X systems and documented support for the SSH protocol.

Paralellization of the MRI simulator slightly differs for motionless (stationary tissue) and moving (blood) object components. In the first case imaged 3D volume is devided into slices along the x dimension. Each node of the grid performs signal acquisition for the whole 3D k-space, but it takes into account only the contribution coming from the object voxels belonging to this nodes slice. Then, the root node collects the k-space volumes from all the nodes and they are aggregated.

After completing calculations for one component it moves on to the remaining ones.

Because, the object component corresponding to blood flowing through vessels is designed differently, it ecquires separate parallelization model. Instead of deviding the object into slices, every node receives a portion of cells to handle. A node is thus responsible for updating cell position and recalculation of its corresponding magnetization vector. During acquistion step, the k-space is filled by summing up signals coming from object cells managed by a particular node.

It is worth to stress that this strightforward parallelization model further substantiates the assumed approach to moving object component defintion. Each object cell constitutes an autonomous agent and all information connected with its replacement and magnetic vector state can be derived within a grid node. If it were the alternative approach based on subvoxel analysis, information about magnetization of flowing liquid would have to be somehow transmitted to neighboring voxels belonging to different nodes slices. This would involve additional communication between working nodes and inevitable transmission synchronization issues leading to downgraded performace.

C. Runtime parameters

In order to launch jobs on SSH-based grid using Open MPI software, it is important to properly setup user accounts on all machines. First of all, a user of the same login name should be registered on every node. Preferably, this user should be granted administrative priviliges. User logging between computers should proceed without password authentication, as described in the Open MPI documentation [16]. To achieve this, one has to generate RSA key pair on the root node, copy the public key to the *authorized_key* file, and eventually distribute *./ssh* directory to the specified user home directories on all remote nodes.

Moreover, the Open MPI distribtion – the same on all nodes at least up to the major subversion number – should be installed at the identical location. Similarily, the MPI program has to placed under the same absolute path on every computer. The progam itself, though uniformly named too, should be architecture and OS specific. Thus, in our grid the same executable is shared between both iMacs and Mac Mini units, but there are separate compilations for Xserve and the Linux server.

To simplify file exchange between nodes, it is possible to establish common file system for them. For this purpose, we utilized the NFS (Network File System) protocol. On the root node one has to create a shared folder and place its path in the */etc/exports* file with appropriately set network address space. Then, on the remote hosts, this shared resource needs to be mounted using *mount -t nfs* shell command.

Eventually, the command which invokes a parallel job using MPI and on the SSH-based grid takes the following form:

```
sh$ $MPI_HOME/mpirun
    --mca plm_rsh_agent ssh
    --hostfile $JOB_HOME/hosts
    --np 50
    $JOB_HOME/simra_mpi
    $NFS_COMMON/job_paramaters_file
```

where `$MPI_HOME` indicates the installation directory of the MPI distribution, shared `$NFS_COMMON` directory contains any addditional files required by the executed program (object definition, flow information, MRI sequence parameters) and `$JOB_HOME` is the path to the parallelized executable. The `--mca` switch states that it is the SSH agent that controls the grid management, the hosts file contains the list of computers in the grid, and the `--np` switch specifies the number of processes to run in parallel.

V. Evaluation of Grid Performance

In order to assess efficiency of the designed computer grid, we have conducted a series of simulations for one digital phantom at different scales of the grid. For every scale, there was a part of CPU cores constantly present across various configurations (Xserve, Linux server, and iMacs) and the rest (Mac Mini computers) were accumulatively added in subsequent simulations.

The phantom used in the experiments is composed of two virtual tissues whose voxels are spherically distributed arround common origin. The resolution of the object is 100 voxels in each dimension, while the physical size were set to $(0.2m)^3$.The intesity levels of the object components range from 0 to 125. These limits results from the assumed method of phantom synthesis. First, one chooses the minimum and maximum radii of each component boundary spheres. Secondly, a voxel is devided into subvoxels, 5 in each direction, giving 125 subvoxels in total. If a center of a subvoxels lies within the assumed limits, then it is added to the object component. The number of subvoxels included determines intensity of its corresponding voxel.

Table I summerizes the details of the object and its components. Note, that T_1, T_2 and ρ parameters are set to values typical to fat and water ((components no. 1 and 2 respectively). Average simulation times achieved to synthesize MR images are shown in Table II, while synthesized images (GE, T_2*-weighted) are presented in Fig. 2. Note that relatively strong T_2* weighting of the image lead to almost entire supression of the signal related to the fat component. On the contrary, water with high T_2 constant appears bright on the image.

VI. Conclusion

Analysis of the obtained time measurements (see Table II) proves efficiency of the assembled computer grid. Also, grid management and inter-process communication controlled by Unix/Linux machines guarantees reliable and stable operation. Computational speedup scales (almost) linearly with the number of CPU cores. Time needed to synthesize high-resolution volumetric images ranges on average from 1.5 up to 2 hours per component, depending on the object size. This appears to be acceptable, especially if compared to the results achieved by other simulators. This leads to conclusion that our

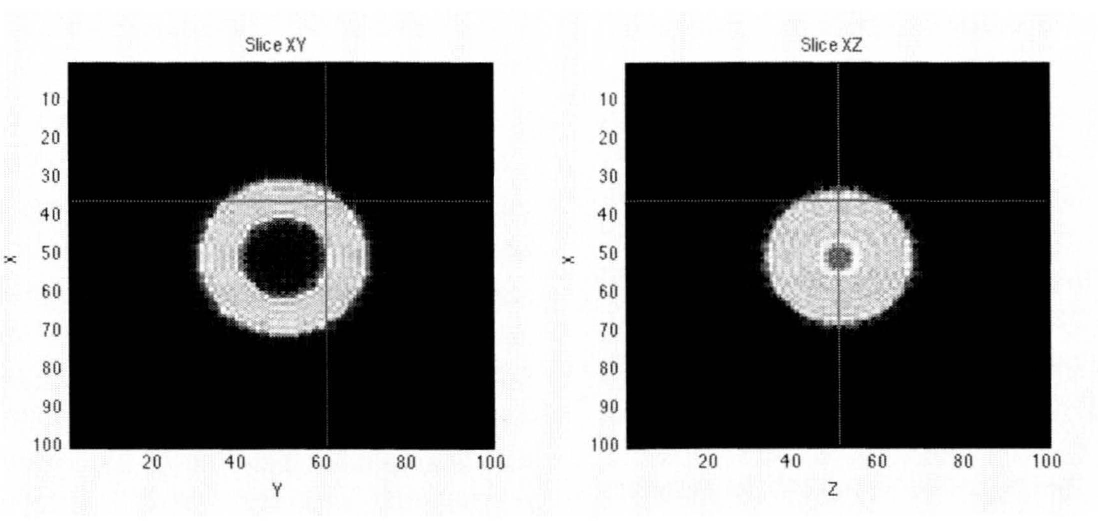


Fig. 2. Synthesized MR image. Gradient echo sequence, TE=15 ms, TR=120 ms, flip angle = 20°

TABLE I
DIGITAL PHANTOM DESCRIPTION

Component No.	$r_{\min}{}^1$	$r_{\max}{}^2$	ρ^3	T_1	T_2
1	2 cm	4 cm	1	350	70
2	4 cm	6 cm	1	2569	329

[1, 2] minimal and maximal sphere radii

[3] water relative proton density

TABLE II
GRID PERFORMANCE EVALUATION

Number of cores	26	30	34	38	42	46	48
Time $\times 10^3 [s]$	12,9	12,3	12,0	11,4	11,1	10,7	10,2

gird can be expected to produce angiographic MR images in reasonable times. However, if the number of available cores occurs insufficient, the archicture of the grid allows its further expansion on additional units.

ACKNOWLEDGMENT

This work was supported by the Polish National Science Centre grant no. 6509/B/T02/2011/40.

REFERENCES

[1] R. Kwan, A. Evans, and G. Pike, "An extensible MRI simulator for post-processing evaluation," in *Visualization in Biomedical Computing*, ser. Lecture Notes in Computer Science, K. Hoehne and R. Kikinis, Eds. Springer Berlin / Heidelberg, 1996, vol. 1131, pp. 135–140. [Online]. Available: http://dx.doi.org/10.1007/BFb0046947

[2] A. Materka, M. Strzelecki, P. Szczypinski, M. Kocinski, A. Deistung, and J. Reichenbach, "Arteries tracking in simultaneous TOF-SWI MR images: image characteristics and preliminary results," in *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*, 2009, pp. 748 –753.

[3] M. Strzelecki, P. Szczypiński, A. Materka, M. Kociński, and A. Sankowski, "Level-set segmentation of noisy 3D images of numerically simulated blood vessels and vascular trees," in *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*, 2009, pp. 742 –747.

[4] D. A. Yoder, Y. Zhao, C. B. Paschal, and J. M. Fitzpatric, "MRI simulator with object-specific field map calculations," *Magnetic Resonance Imaging*, vol. 22, pp. 315–328, 2004.

[5] H. Benoit-Cattin, G. Collowet, B. Belaroussi, H. Saint-Jalmes, and C. Odet, "The SIMRI project: a versatile and interactive MRI simulator," *Journal of Magnetic Resonance*, vol. 173, pp. 97–115, 2005.

[6] T. Stoecker, K. Vahedipour, and N. J. Shah, *HPC Simulation of Magnetic Resonance Imaging*, ser. NIC Series. Juelich: John von Neumann Institute for Computing, 2007, vol. 38, pp. 155–164.

[7] K. Jurczuk and M. Kretowski, "Virtual magnetic resonance imaging – parallel implementation in a cluster computing environment," *Biocybernetics and Biomedical Engineering*, vol. 29, no. 3, pp. 31–46, 2009.

[8] A. Simmons, S. Arridge, G. Barker, and S. Williams, "Simulation of MRI cluster plots and ap," *Magnetic Resonance Imaging*, vol. 14, no. 1, pp. 73–92, 1996.

[9] J. S. Petersson, J. O. Christoffersson, and K. Golman, "MRI simulation using the k-space formalism," *Magnetic Resonance Imaging*, vol. 11, no. 4, pp. 557–568, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0730725X9390475S

[10] Z.-P. Liang and P. C. Lauterbur, *Principles of Magnetic Resonance Imaging*, ser. Series in Biomedical Engineering. IEEE Press, 2000.

[11] P. Tofts, *Quantitative MRI of the Brain: measuring changes caused by desease*. Chichester: John Wiley & Sons, 2003.

[12] MATLAB, *version 7.13.0 (R2011b)*. Natick, Massachusetts: The MathWorks Inc., 2011.

[13] C. Multiphysics, *version 4.2.1 (4.2a)*. Los Angeles, CA: COMSOL, Inc., 2012.

[14] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Applications*. Springer-Verlag, 2008.

[15] N. Karonis, B. Toonen, and I. Foster, "MPICH-G2: A grid-enabled implementation of the Message Passing Interface," *Journal of Parallel and Distributed Computing*, vol. 63, no. 5, pp. 551–563, 2003.

[16] E. Gabriel, G. E. Fagg, G. Bosilca, T. Angskun, J. J. Dongarra, J. M. Squyres, V. Sahay, P. Kambadur, B. Barrett, A. Lumsdaine, R. H. Castain, D. J. Daniel, R. L., and T. S. Woodall, "Open MPI: Goals, concept, and design of a next generation MPI implementation," in *Proceedings, 11th European PVM/MPI Users' Group Meeting*, Budapest, Hungary, 2004, pp. 97–104.

NTAV/SPA 2012
27-29th September, 2012, Łódź, Poland
New trends in audio and video / Signal processing algorithms, architectures, arrangements and applications

Mathematical Morphology Analysis of 3D MRA Images of Human Brain for Estimation of Blood Vessels Parameters

Adam Sankowski, Andrzej Materka
Lodz University of Technology
Institute of Electronics
Lodz, Poland
sankowski@gmail.com

Abstract — **The aim of the described research is to develop brain blood vessels modeling algorithm on the basis of three-dimensional high-resolution magnetic resonance images. The geometrical model of vessels will be used to visualize arteries and veins and to determine their quantitative description. It is expected that the estimated geometrical parameters of vessels will significantly complement the results of medical examinations and contribute to a more objective and accurate medical diagnosis. Geometrical parameters need to be extracted from MRA image after segmentation. Most important parameters are: vessel diameter, local direction and location of vessel endings and bifurcations. In this paper, the estimation algorithms will be presented. They are developed on the basis of three-dimensional skeleton of vessels and vessel tracking techniques.**

Keywords — ***Magnetic Resonanse Angiography; blood vessels; image processing; parameter estimation; diameter; local direction; bifurcation***

I. Introduction

A. Problem description

Development of magnetic resonance technology enables one to acquire high resolution, three dimensional images. Time of Flight [1] and Susceptibility Weighted Imaging [2] techniques combined together in one measurement result in a full map of veins and arteries [3]. Typically, images have to be segmented for vessel extraction. There are many approaches to segmentation [4]. They are divided to two main groups: based on geometrical modeling and on mathematical morphology. This paper will focus on the processing step performed after segmentation witch is modeling. To produce reliable geometrical model of vessels one need an information about vessel tree parameters. Those parameters can be estimated from segmented MR images. Assumption is that estimation starts from binary, three dimensional image where all vessel voxels have brightness value equal 1 and background equals 0 (Fig. 1a). Another necessary input is skeleton of vessels which will be treated as center line of vessels (Fig. 1b). Morphological thinning has been chosen from a range of different skeleton methods. Morphological thinning is most suitable for vessel tracking purposes because of maximum reduction of vessel region. With this approach each voxel in the center line of vessel will have only 2 neighbor voxels which also are part of the center line in all-26-voxel neighborhood. Exceptions are bifurcation voxels (3 neighbors) and ending voxels (1 neighbor).

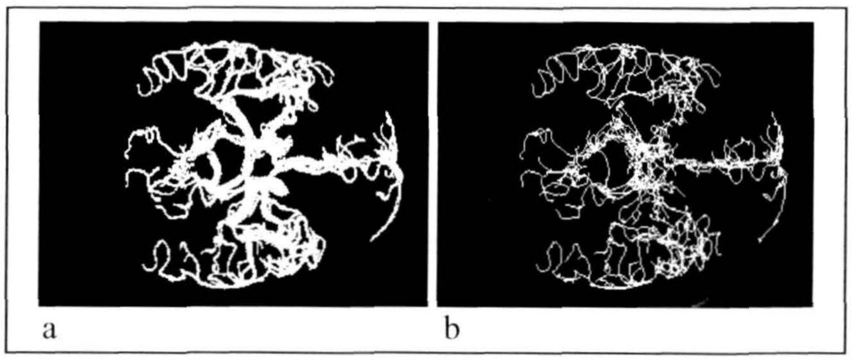


Figure 1. a)Maximum intensity projection of segmented 3D MRA image, b) maximum intensity projection of the result of thinning applied to image (a)

The most important vessel parameters, considered in this paper, are: vessel diameter, local direction as well as locations of vessel endings and bifurcations (Fig. 2). Estimation of vessel diameter can be performed by using both of those images.

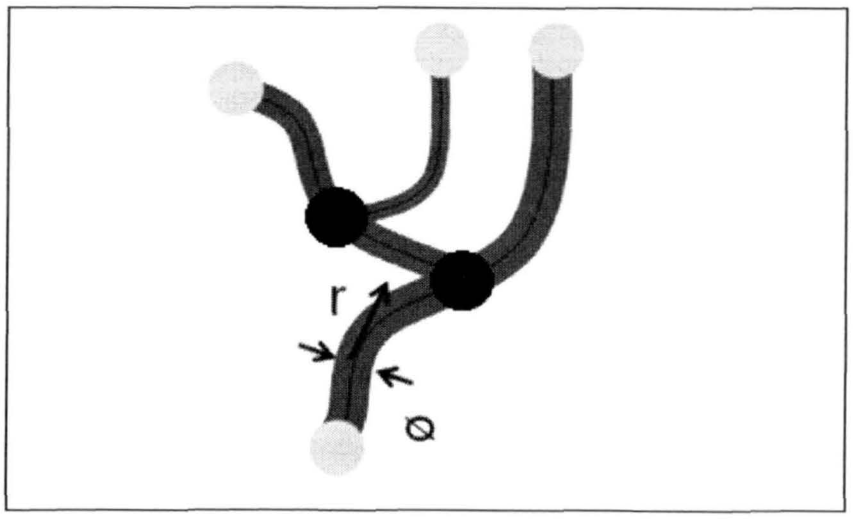


Figure 2. Parameters of vessel tree: diameter, local direction, location of endings and bifurcation

B. Methods overview

There are many approaches of diameter estimation. Majority of them are using skeleton to obtain center line of vessel. Method proposed by Sorantin [5] is based on user interaction. Shortest path between two points determined by user is computed. Chillet [6] suggest to use normal plane and cross-section area. That approach can produce diameter information for every skeleton voxel. Another method, employing deformable sphere located at the center line, is proposed by Zhou [7]. Method proposed in this paper is similar, but based on binary ball structuring element. The center of the ball is placed at every voxel of center line, but in vessels image. Ball starts from smallest possible radius. When ball contains vessel voxels only, its radius is increased. This operation is repeated until some part of "growing" ball will contain background voxels. Local vessel diameter is computed based on number of iterations and percent of vessel voxels in last iteration. Tests were performed on digital tube phantoms with different position and diameter. The aim of the testing was to assess the accuracy of the estimated diameter.

Another important geometrical parameters are locations of bifurcations and endings of vessels. Bifurcation detection in vessel tree are usually designed for two dimensional images. To find bifurcation in this space, it is enough to use a set of masks [8] or count neighbors [9]. AdaBoost method of Gaussian filter and first and second derivatives of Gaussian filter[7] are used for three-dimensional images. Method proposed in this paper is based on fingerprint line tracking[10], but performed in three dimensional space. Only center line image is used to find locations of bifurcation and endings. Ending voxels are easy to find. Endings in center line image received from morphological thinning have only 1 neighbor. Voxels detected as endings are pending starting points for vessel tracking. The 26 voxel neighborhood is checked for center line voxels. When found it becomes next voxel to check. Traveled path becomes invisible for later tracking. When tracking finds more than one path to follow, the path is chosen randomly and start point of alternative path is added to an endings search queue. When tracking ends in a voxel not labeled as ending, then that voxel becomes labeled as bifurcation. In vessel trees with more bifurcations this is not enough. It may be necessary to search for endings in whole image after tracking and making traveled path invisible. Found endings will also be marked as bifurcations.

Finally, local direction of vessels is also an important parameter to find. Estimation method is in this case similar to the previous one. Vessel tracking is performed on center line image, but requires information about all previously estimated parameters (diameter, bifurcations and endings). Tracking starts from voxel which is labeled as vessel ending and has largest diameter. Tracking makes traveled path invisible like in bifurcation search. Local direction is determined in every voxel on centerline except for bifurcations and endings. When previous voxel in path is invisible, then examined voxel have only one neighbor. Location of that neighbor relative to examined voxel gives information about local direction. When tracking finds bifurcation it travels random branch. Beginning of second branch is added to queue as another starting point. Points from queue are used to start new tracking when previous one ends (when voxel labeled as vessel ending is found).

II. Parameter Estimation

A. Estimation of diameter

Diameter estimation is performed in every voxel of vessel center line. Voxel index is taken from center line image and put into full vessel image as center of binary ball. Ball radius is equal to 1 voxel. Inner product of ball is computed. Brightness value of every vessel voxel equals 1 so inner product value is equal to number of vessel voxels inside ball (Fig. 3). When every voxel within given ball radius equals 1, then radius is increased. This process is repeated until the ball is fully filled with vessel voxels. Information about biggest ball radius without background voxels and percent of vessel voxels in larger radius is acquired by performing this procedure for every voxel of the center line.

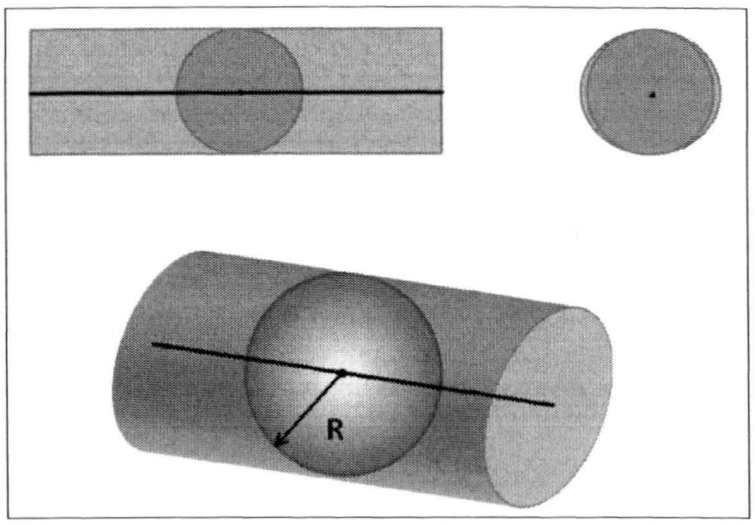


Figure 3. Visualization of diameter estimation

This information needs proper interpretation. This was done by using digital phantoms in tube shape. Tubes represents vessels with different radius (from 1 to 10 voxels) and different directions in three-dimensional space. Nearly 200 phantoms (Fig. 4) with known diameter were examined by diameter estimation algorithm.

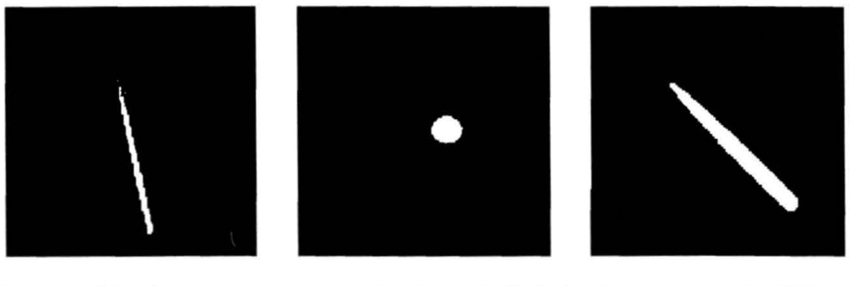


Figure 4.Maximum intensity projection of digital phantoms with different diameters and different orientation

Information about biggest radius of the ball fully filled with vessel voxels and information about percentage of ball voxels located inside the vessel at increased ball radius were compared with information about real radius of phantoms. Aim of composition was to reduce difference between real diameter and estimated diameter. After error reduction estimated diameter is over 90% correct. Usually mistakes are not greater than 1 voxel. Example results are shown on Fig. 5.

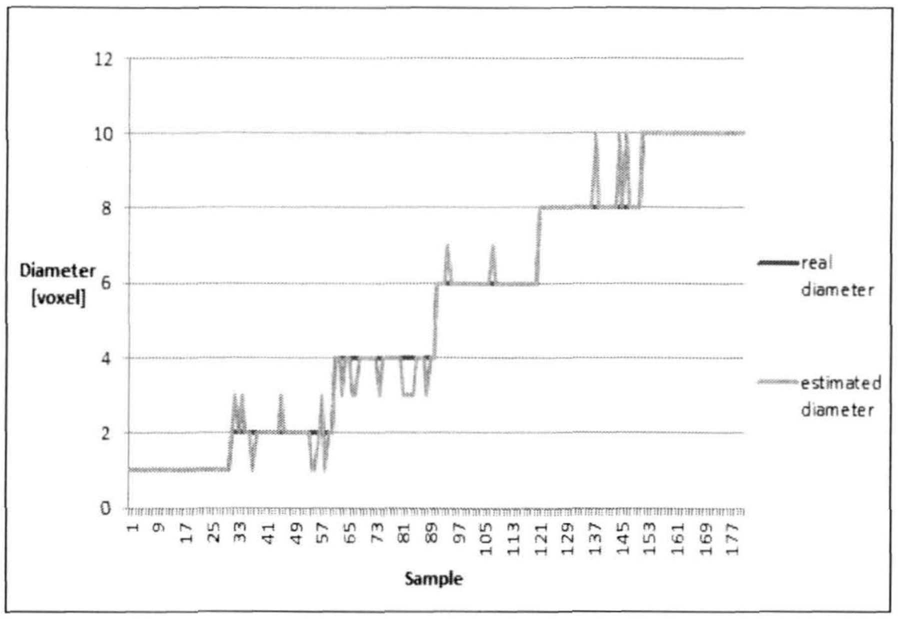


Figure. 5 Comparison of diameter estimated by the proposed method with correct diameter

B. Localization of bifurcations and ending

The algorithm is based on center line image only. Image is produced by morphological thinning. The algorithm returns another centerline image but with changed brightness values. Brightness value equal 3 means that a voxel has been labeled as vessel ending. Brightness value equal 4 means that voxel is labeled as point of bifurcation.

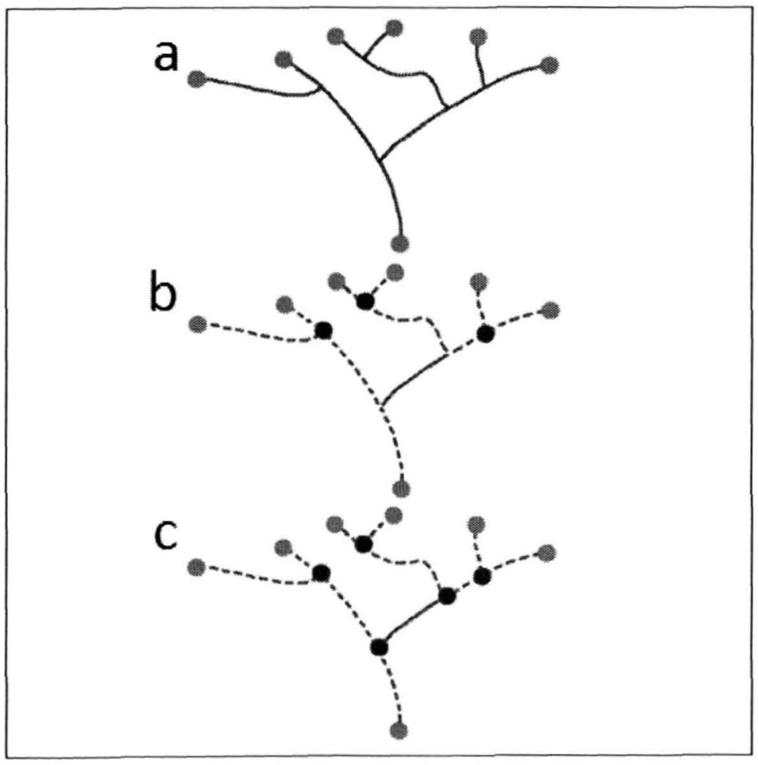


Figure 6. Bifurcations and endings estimation steps

First task (Fig. 6a) in the proposed algorithm is to find endings of vessels. Thanks to morphological thinning, the ending voxels have only one neighbor in full 26-neighborhood. Brightness value of those voxels is set to 3. Second step (Fig. 6b) is vessel tracking. Purpose of tracking is to find bifurcation voxels. All indexes of ending voxels are put into a queue as start voxels for tracking. Tracking changes brightness value of current voxel to 2 if current voxel isn't labeled as ending or bifurcation. Thanks to that traveled path is labeled and other tracking won't follow that path. Neighborhood of currently examined voxel checking is performed in every step of tracking. Usually only one unlabeled neighbor is found. In that case, a found neighbor becomes next voxel to check. When tracking arrive to crossroads (two unlabeled neighbors found in neighborhood) further path is chosen randomly. Third case is end of road (no unlabeled neighbors found). In this case it can be another ending voxel. If it is (current voxel brightness value equals 3), then tracking ends and another tracking starts from next start point from queue. If current voxel isn't labeled as ending it means that this is bifurcation or voxel next to bifurcation. Current voxel is labeled as bifurcation. It may happen that not all bifurcations are found. Because of that third step (Fig. 5c) repeats searching of ending voxels, but operates only on unlabeled voxels. Found voxels are labeled as bifurcations.

There is one issue with the algorithm. Sometimes voxel labeled as bifurcation isn't really a bifurcation, but some neighbor is. This issue requires correction which is realized by simple loop. If voxel labeled as bifurcation has only two neighbors, neighborhood of that voxel is searched to find voxel with tree neighbors. If such neighbor was found, then it becomes labeled as bifurcation and voxel previously labeled as bifurcation becomes part of center line. This procedure is enough to detect all bifurcations in real MRA image. Algorithm testing on more complicated digital vessel tree phantoms showed that when number of bifurcations in vessel tree is large then not all of them are detected. Because of that estimation algorithm became iterative and is repeated on unlabeled voxels until there are no unlabeled voxels left in entire image. Thanks to that, localization of bifurcations works well regardless of vessel tree complexity.

C. Estimation of local direction

Local direction estimation results in new image of vessel center line with changed brightness values just like in bifurcation estimation. In this image brightness value represents local direction in every center line voxel except for endings and bifurcations. Spectrum of those values is from 1 to 26 just like number of neighbors in three dimensional space. Information about previously estimated parameters like diameter, endings and bifurcations is required to create such image. Local direction is estimated through tracking similar to that used is bifurcation estimation. Tracking starts from voxel with biggest diameter which is labeled as ending. Traveled path is labeled to make it invisible. Examined voxel usually has one unlabeled neighbor. Position of that neighbor in one of 26 directions is written as brightness value in output image and found neighbor is set as next voxel to examine. If currently examined voxel is labeled as bifurcation then randomly chosen unlabeled voxel is set as next voxel to check and index of other is put into a queue. When tracking arrive in voxel labeled as ending, then next voxel index is taken from queue and new tracking starts. When queue becomes empty and some unlabeled voxels are still in image, whole process is repeated only on unlabeled voxels. This is necessary because in real MRA image of brain there are two main arteries without any connection.

III. FUTURE WORK

Proposed solutions need further improvements. Bifurcation and ending estimation still generates minor mistakes and must be reviewed. Another important modification will be diameter estimation performed on gray scale images by computing standard deviation inside ball radius. This will allow to express diameter as real number instead of natural number. When all parameters will be estimated correctly then vessel modeling can be started.

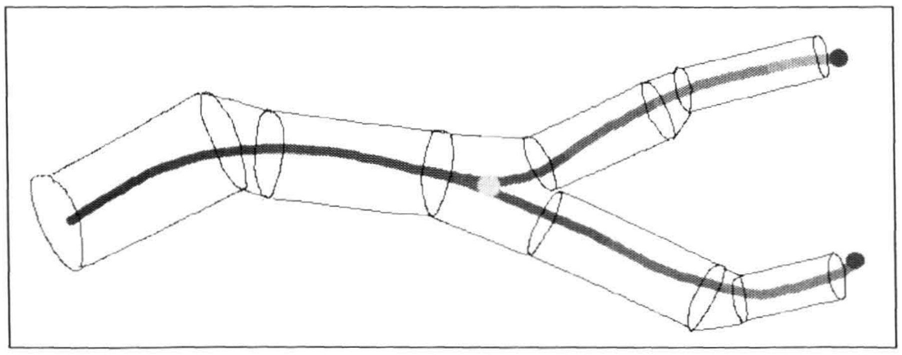


Figure 7. Model based on cylinders and trapezoids

Model will be based on cylinders and trapezoids like in Fig. 7. All gathered information about vessel parameters will be useful while building model. Reliable 3D model of real brain vessels can be used for blood flow simulations

REFERENCES

[1] Bernstein M. A.,. Huston J, Lin C.,Gibbs G. F., Felmlee J. P., High-resolution intracranial and cervical MRA at 3.0T: technical considerations and initial experience, Magn Reson Med, 46 (2001), no.5, 955-962.

[2] Reichenbach J.R., Haacke E.M.: High resolution BOLD venographic imaging: A window into brain function, NMR Biomed, 14 (2001), 7-8, 453-467.

[3] Deistung A., Dittrich E., Sedlacik J., Rauscher A., Reichenbach J., ToF-SWI Simultaneous time of flight and fully flow compensated susceptibility weighted imaging, Journal of Magnetic Resonance Imaging, 29 (2009), no.6, 1478-1484

[4] Kirbas C., Quek F., A review of vessel extraction techniques and algorithms, ACM Computing Surveys, 36 (2004), no.2, 81-121

[5] E. Sorantin, C. Halmai, B. Erdohelyi, K. Palagyi, L.G. Nyul, K. Olle, B. Geiger, F. Lindbichler, G. Friedrich, K. Kiesler, Spiral-CT-based assessment of tracheal stenoses using 3D skeletonization. IEEE Transactions on Medical Imaging 21 (2002) 263–273

[6] D. Chillet, N. Passat, M.-A. Jacob-Da Col, and J. Baruthio, Thickness Estimation of Discrete Tree-Like Tubular Objects: Application to Vessel Quantification, Proceeding SCIA'05 Proceedings of the 14th Scandinavian conference on Image Analysis Pages 263-271

[7] Jinghao Zhou, Sukmoon Chang, Dimitris Metaxas and Gig Mageras, 3D-3D Tubular Organ Registration and Bifurcation Detection from CT Images, Conf Proc IEEE Eng Med Biol Soc. 2008;2008:5394-7

[8] Alauddin Bhuiyan, Baikunth Nath, Joselito Chua and Kotagiri Ramamohanarao, Automatic detection of vascular bifurcations and crossovers from color retinal fundus images, Proceeding SITIS '07 Proceedings of the 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System Pages 711-718

[9] Edoardo Ardizzone, Roberto Pirrone, Orazio Gambino and Salvatore Radosta, Blood Vessels and Feature Points Detection on Retinal Images, Conf Proc IEEE Eng Med Biol Soc. 2008;2008:2246-9.

[10] Juan Erez , Amengual , A Juan , J C P'erez , F Prat , J M Vilar, Real-Time Minutiae Extraction In Fingerprint Images, IPA97, 15- 17 July 1997, Conference Publication No. 443 0 IEE

Comparison of ASM and AAM-based Segmentation of Prostate Image in the CT Scans for Radiotherapy Planning

Artur Kos, Tomasz Zieliński
AGH University of Science and Technology
Department of Telecommunications
Al. Mickiewicza 30, PL30059, Krakow, Poland
kosar@agh.edu.pl, tzielin@agh.edu.pl

Paweł Kukolowicz
The Jan Kochanowski University
of Humanities and Sciences in Kielce
Department of Medical Physics
Świętokrzyska 15, PL25406, Kielce, Poland

Andrzej Skalski
AGH University of Science and Technology
Department of Measurement and Electronics
Al. Mickiewicza 30, PL30059, Krakow, Poland
skalski@agh.edu.pl

Piotr Kedzierawski, Tomasz Kuszewski
HollyCross Cancer Center
Department of Radiotherapy
Artwińskiego 3, PL25734, Kielce, Poland
piotr.kc@op.pl

Abstract — **In this paper we present a novel method of medical CT data segmentation using explicit atlas-type knowledge and Active Appearance Model (AAM) as a principal segmentation tool. New approach of automatic landmarks creation for AAM is proposed. Obtained results of automatic segmentation of prostate image in the CT scans are compared with true outlines drawn by medical doctors during the radiotherapy planning and with the Active Shape Model (ASM)-based segmentation tested by us before. Using ASM turned out to be better than AAM.**

Keywords — ***ASM, AAM, CT, segmentation, prostate radiotherapy, atlas segmentation***

I. Introduction

High numerical power of contemporary computers makes available application of many different advanced algorithms supporting biomedical diagnosis and therapy. In particular, thanks to this, complex procedures of biomedical imaging can be much easier, wider and faster exploited at present. It is especially important in medical radiology where X-ray technique is a standard tool used for, both, medical imaging/diagnosis of cancer changes (with low radiation dose) and their radiotherapy/treatment (with high radiation dose). In both tasks the computed tomography is used and its functionalities (reconstruction speed, accuracy, etc.) should be improved together with associated data analysis and processing (segmentation, registration, etc.).

A prostate cancer is one of the most common cancer in men. Age and genetic predisposition are its main risk factors. Choosing medical treatment depends on the disease advance, patient age and co-existing illnesses. In radical treatment of locally advanced cancer, a surgical treatment is preferred for younger persons while for older ones – the radiotherapy.

Before the radiotherapy the computed tomography (CT) examination/imaging is done and resulting data are used for preparation of *treatment plans*. It is necessary to delineate the cancer tissue in the transverse CT images (i.e. to mark the GTV – *Gross Tumor Volume*) as well as to contour anatomical structures that should be particularly protected against radiation. In most clinical cases a region to be radiated, attacked by the cancer, consists of the prostate and seminal vesicles while rectum, bladder and heads of femurs (tight bones) should be not radiated.

In this paper, a problem of GTV segmentation is presented. The main problem in the prostate radiotherapy is a low quality of the CT data causing that indicating the border between such anatomical structures like bladder and prostate is very difficult. Since a priori information can significantly help interpret new data that are analysed, nowadays segmentation algorithms are supported by an additional explicit knowledge, very often in the form of medical models and atlases, and work much better than simple unsupervised solutions. Segmentation of a new CT can be guided by a pixel-type or a contour-type information extracted from available CT (the same modality) [1,2] or MRI (different modality in which soft tissue is better visible) [3,4] databases. Active shape models (ASM) [5] and active appearance models (AAM) [6,7] are knowledge extraction (approximation) methods used for building statistical generalizations of collected data coming from different realizations of the same or similar objects, e.g. human faces. The AAMs have been already widely applied in 3D medical image segmentation [8]. The main problem making their application difficult is required consistent (strict correspondent) medical landmarks placement over a large database (e.g. medical annotation of the same points of the same human organ in many CTs of different patients). Some solutions to this problem has been already proposed and applied [9,10] but the problem is still opened.

In this paper we address a problem of segmentation of the prostate image in CT scans, supervised (aided) by explicit statistical active appearance model of the prostate build by AAM (mean value and its perpendicular deviations) in similar way as in [11] where the ASM was exploited. Consistent placement of prostate landmarks in many CTs, required by the AAM, has been solved in a novel way. All CTs have been registered in a groupwise manner using the free form deformation method with B-splines (global and local voxel adjustment) [12] what allows, in conjunction with interpolation and using inverse deformation field with geometrical correction, to find sets of strictly corresponding landmarks. These sets were used by AAM to a statistical prostate shape calculation. Manual landmarks' sets generation on the mean CT is also possible.

The paper consists of introduction, brief state-of-the art problem description, presentation of the proposed method (training data generation, statistical model of prostate image construction and AAM-based segmentation), presentation of obtained results in comparison with ASM methodology used in [11] and their discussion.

II. Proposed Method and Methodology

The main problem existing in application of the 3D AAM technique is appropriate creation of corresponding (consistent) landmark sets of training data. It is a direct consequence of the requirement that each landmark marked in one training data should correctly correspond to the same point in other training data. This feature is very difficult to obtain from practical point of view since medical doctors mark arbitrarily sets of landmarks and the landmark consistence (correspondence) should be obtained in large database, especially in 3D. The same problem exist for the ASM that was exploited in our previous work [11]. Otherwise an incorrect parametrisation of the object would result. Another solution is to generate correct sets of landmarks automatically. Frangi et al. [13,14] proposed registration of an analysed data to an available atlas using quasi-affine rigid global transformation and local non-rigid elastic transformation. Landmarks themselves are put in the atlas data, copied to the registered analysed data and then projected back to the atlas aligned-coordinates using inverse of the known elastic transformation. Next, AAM can be performed on consistent sets of landmarks associated with data of interest. The other solution is to realize jointly a rigid global and an elastic local transformation [15,16] or to re-sample all training data into the same number of slices by interpolating elliptic Fourier descriptor (EFD) coefficients [17].

In this paper a segmentation of the prostate image in the CT scan using the AAM technique [6,7,18] in connection with group wise registration [12] is proposed. Such approach ensures automatic landmark creation. Its differences in respect to [13,15] and [1] were stated in the introduction. In the first part of the research the prostate statistical model was built. Available CT images of prostates of different patients were registered to one arbitrary chosen co-ordinate system in voxel-to-voxel manner, i.e. each CT image was processed by the global affine and local B-spline transformations (FDD method) [12]. Using found values of transformations' parameters prostate contours annotated (drawn) by medical doctors were transformed from the input CTs to the common co-ordinate system, then averaged and interpolated (re-sampled). The resultant approximation was next transformed back to each input CT. Then, a geometrical correction to the medical contours was done followed by the ICP (Iterative Closest Point) method. This way the set of CTs with corresponding prostate landmarks was obtained. Its possession is absolutely necessary on input of active shape modeling (ASM) procedures but usually it is very difficult to get. Finally, the AAM of the prostate based on the Kroon's implementation [19] was built, namely: the prostate statistical shape, texture spread of its points and their variances in perpendicular direction to the surface correlated with the texture. In the second part of the research, an exemplary automatic segmentation of prostate images for 9 validation CT sets was done. Results obtained for the AAM method were compared with the similar ASM application [11] for the same test set.

III. Training Data Generation

The idea of the applied training data generation for the AAM is presented in Fig. 1. First, the automatic groupwise registration is performed, i.e. each image is registered to one coordinate system and its deformation field $\mathbf{T}_i$ in respect to this system is found. The coordinate system (the reference) is not defined explicitly, but is calculated implicitly by constraining

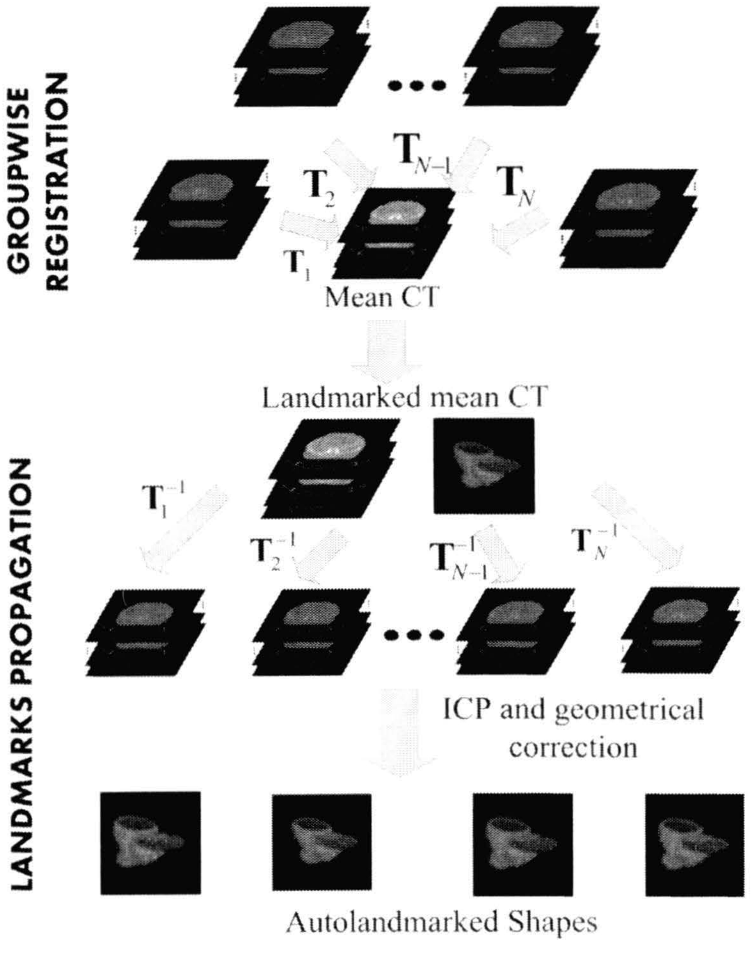


Figure 1. Block diagram of landmarks of training data generation. Description in the text

the average deformation to be the identity transform [12]. In the groupwise registration framework, we have chosen the B-spline Free Form Deformation (FFD) working in 3-level multiresolution scheme, proposed by Rueckel et al. [20] to registration of MRI breast images. The non-rigid B-Spline FFD was combined with an affine transformation:

$$\mathbf{T}(x,y,z) = \mathbf{T}_{\mathrm{local}}(\mathbf{T}_{\mathrm{global}}(x,y,z)) \qquad (1)$$

where $\mathbf{T}_{\mathrm{global}}$ is the affine transform and $\mathbf{T}_{\mathrm{local}}$ is the deformation model based on B-splines. The affine transform allows to compensate a patients body pose while the B-splines – a local deformation between the data. Calculation of deformation fields was done using implementation given by Balci et al. [12].

Then, a set of M $\widehat{\mathbf{p}} = [\hat{x}_1, \hat{y}_1\hat{z}_1, \dots, \hat{x}_M, \hat{y}_M\hat{z}_M]$ landmarks on gradient of the mean CT $\hat{\mathbf{I}}$ was calculated (i-th are landmarks coordinates). Knowing the N forward transformations of N CT data sets we can calculate the inverse (backward) transformations $\mathbf{T}_i^{-1}$ ($i = 1..N$) what allows to map landmarks from the mean CT to individual data used to the mean CT creation. Resulting positions were corrected using the ICP method and a distance to the doctors contour. Finally, we received N vectors of landmarks correlated with the training data with correspondences between them kept:

$$\mathbf{p_k} = \mathbf{T}_i^{-1}[\hat{x}_1, \hat{y}_1\hat{z}_1, \dots, \hat{x}_M, \hat{y}_M\hat{z}_M] \qquad (2)$$

IV. MODEL CONSTRUCTION

Creation of the statistical model of the prostate image is based on the Active Appearance Model algorithm introduced by Cootes et al. [20,22]. During its construction the following quantities were computed:

- mean shape $\overline{\mathbf{P}}$ of landmarks $\mathbf{p}_i$:

$$\overline{\mathbf{P}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{p}_i \qquad (3)$$

$$\mathbf{p}_i = \mathbf{T}_i^{-1}[\hat{x}_1, \hat{y}_1\hat{z}_1, \dots, \hat{x}_M, \hat{y}_M\hat{z}_M] \qquad (4)$$

- covariance matrix S_b:

$$S_b = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{p}_i - \overline{\mathbf{P}})(\mathbf{p}_i - \overline{\mathbf{P}})^T \qquad (5)$$

- eigenvectors $\mathbf{\Phi}_{si}$ of S_b (using principal component analysis (PCA)):

$$S_b\mathbf{\Phi}_{si} = \lambda_i\mathbf{\Phi}_{si} \qquad (6)$$

$$\mathbf{\Phi}_s = [\mathbf{\Phi}_{s2}, \mathbf{\Phi}_{s2}, \dots \mathbf{\Phi}_{s2M}] \qquad (7)$$

- vector $\mathbf{b}_s$ such that for any $\boldsymbol{\pi}$ the following equation holds:

$$\boldsymbol{\pi} = \overline{\mathbf{P}} + \mathbf{\Phi}_s\mathbf{b}_s, \quad \mathbf{b}_s = [\mathrm{b}_{s1}, \mathrm{b}_{s2}, \dots \mathrm{b}_{s2M}]^T \qquad (8)$$

The vectors $\mathbf{\Phi}_s$ and $\mathbf{b}_s$ represent, respectively, main directions of *mean* model changes and variances along these directions. Since eigenvectors with higher indexes have smaller deviations, only the t largest eigenvalues were retained which allows to reduce dimensionality of the model. The final shape model was described by:

$$\boldsymbol{\pi} \approx \overline{\mathbf{P}} + \mathbf{\Phi}_{\mathrm{st}}\mathbf{b}_{\mathrm{st}} \qquad (9)$$

As most shapes are within the range ±3σ, knowing vectors $\mathbf{\Phi}_{\mathrm{st}}$ and $\mathbf{b}_{\mathrm{st}}$ a new shape $\boldsymbol{\pi}$, which is not derived from the training data, can be generated.

As a result of the mentioned above operations, the Active Shape Model (ASM) was obtained. To build the AAM on its base, the following quantities were necessary:

- a vector $\mathbf{g}$ achieved by warping prostate textures of each CT to the points of the mean shape, and normalized using a linear transformation:

$$\mathbf{g} \to (\bar{\mathbf{g}} - \mu_g\mathbf{1})/\sigma_g \qquad (10)$$

  where $\bar{\mathbf{g}}$ is a mean texture, $\mathbf{1}$ is a vector of ones, μ_g and σ_g^2 are the mean and variance of elements of $\mathbf{g}$,

- a texture model computed in the same way as the shape model:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{\Phi}_{\mathrm{gt}}\mathbf{b}_{\mathrm{gt}} \qquad (11)$$

- a vector $\mathbf{c}$ of correlations between the shape and texture, which are learned to generate a combined appearance model in the way described in [22].

Finally, the obtained AAM can be defined by equations:

$$\boldsymbol{\pi} = \overline{\mathbf{P}} + \mathbf{Q}_s\mathbf{c} \qquad (12)$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{Q}_g\mathbf{c} \qquad (13)$$

where $\mathbf{Q}_s$ and $\mathbf{Q}_g$ are matrices describing the modes of variations, derived from the training set.

V. SEGMENTATION USING AAM

To initialize the AAM algorithm the statistical model of the prostate image is approximately fitted to the image data by the global transformation $\mathbf{T}_{\mathrm{global}}$ what gives:

$$\mathbf{\Pi} = \mathbf{T}_{\mathrm{global}}(\overline{\mathbf{P}} + \mathbf{\Phi}_{\mathrm{st}}\mathbf{b}_{\mathrm{st}}) \qquad (14)$$

During matching voxels $\mathbf{g}_{im}$, contained in a region of the image pointed by the vector $\mathbf{\Pi}$, are sampled and projected into the texture model frame $\mathbf{g}_s = \mathbf{T}_{\mathrm{global}}^{-1}(\mathbf{g}_{im})$. As the current model texture is given by $\mathbf{g}_m = \bar{\mathbf{g}} + \mathbf{Q}_g\mathbf{c}$, the error vector (measured in the normalized texture frame and describing differences between a present image and the model) is given by:

$$\mathbf{r}(\mathbf{h}) = \mathbf{g}_s - \mathbf{g}_m \tag{15}$$

where $\mathbf{h} = (\mathbf{c}, \mathrm{s}, \theta, \mathbf{\Pi}_c)$ represents the following parameters of the model: correlation between shape and texture, scale, pose and current landmarks positions.

Let us assume that an RMS of the vector **r** elements is a convergence measure, $\mathrm{E}(\mathbf{h}) = \mathbf{r}^{\mathrm{T}}\mathbf{r}$. Its first order Taylor expansion is equal:

$$\mathbf{r}(\mathbf{h} + \delta\mathbf{h}) = \mathbf{r}(\mathbf{h}) + \frac{\partial \mathbf{r}}{\partial \mathbf{h}}\delta\mathbf{h} \tag{16}$$

and a value of $\delta\mathbf{h}$ which minimizes the formula $|\mathbf{r}(\mathbf{h} + \delta\mathbf{h})|^2$ is to be found. Setting equation (16) to 0 leads to the following solution:

$$\delta\mathbf{h} = -\mathbf{R}\mathbf{r}(\mathbf{h}), \mathbf{R} = \left(\frac{\partial \mathbf{r}^{\mathrm{T}}}{\partial \mathbf{h}} \frac{\partial \mathbf{r}}{\partial \mathbf{h}}\right)^{-1} \frac{\partial \mathbf{r}^{\mathrm{T}}}{\partial \mathbf{h}} \tag{17}$$

The AAM algorithm is started near the target and realizes the following steps [20]:

1. projecting texture sample into the texture model frame by $\mathbf{g}_s = \mathbf{T}^{-1}_{\mathrm{global}}(\mathbf{g}_{im})$;
2. calculating the current error vector $\mathrm{E}(\mathbf{h})$;
3. computing predicted parameters of the displacement vector $\delta\mathbf{h} = -\mathbf{R}\mathbf{r}(\mathbf{h})$;
4. updating the model parameters $\mathbf{h} \rightarrow \mathbf{h} + k\delta\mathbf{h}$, initial value of k is equal 1;
5. computing new points, $\mathbf{\Pi}'$ and the model frame texture $\boldsymbol{g}'_m$;
6. sampling the image at new points in order to get $\boldsymbol{g}'_{im}$;
7. calculating a new error vector $\mathbf{r}' = \mathbf{T}^{-1}_{\mathrm{global}'}(\boldsymbol{g}'_{im})$.

Steps 1-7 are repeated until

8. Changing value of k to 0.5, 0.25 etc. unless $|\mathbf{r}'|^2 < \mathrm{E}$,

then obtained value of the vector **h** allows to compute positions of landmarks what is considered as the final result of matching and treated as the segmentation result.

VI. Results

In order to build an atlas and to train the AAM model we used images coming from 14 patients and the same conventional CT device. As the proposed method finally should find application in positioning patients during the irradiation, for algorithm validation a set of 9 sets of different CT scans performed on Siemens CT-on-Rails were used.

All CT scans were acquired without contrast enhancement. In axial plane the training images had 512 on 512 pixels, while each validation set has different size. Images used for both, validation and training had 1 mm resolution for 5-mm thick slices.

Contours of each patient's prostate, manually drawn by the same medicine doctor, were considered as correct shapes. As a measure of validation accuracy a Dice coefficient has been used.

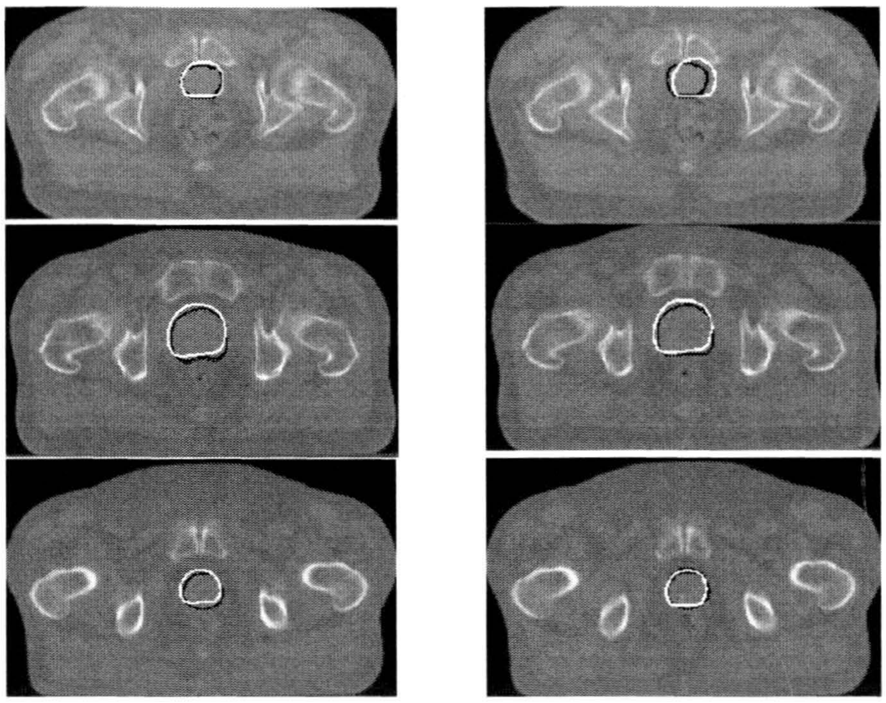


Figure 2. Exemplary results of segmentation using the proposed method with the ASM-based (left column) and AAM-based (right column) algorithm for three different patients. White contour - doctors outline, black contour - results from the proposed method

In [11] there was presented a similar approach to the same problem but making use of the ASM as a segmentation algorithm instead of the AAM. Despite using all image information including textures by the AAM, a low contrast, lack of specific features and similar appearance of the neighbouring organs textures caused that the obtained total accuracy of the AAM-based segmentation algorithm was lower than for the ASM. Using AAM as a segmentation tool led to automatic delineation with mean efficiency specified by Dice coefficient $\boldsymbol{\delta}$=0.652 and with standard deviation $\boldsymbol{\sigma}$=0.0663, while the intensity gradient-oriented ASM allowed to obtain $\boldsymbol{\delta}$=0.812 and $\boldsymbol{\sigma}$=0.045. Both algorithms were tested using *Matlab* on a PC class computer with 2.0 GHz Dual Core processor and 4GB RAM memory. Matching model to the prostate structure on CT scans takes the AAM and ASM about 30 and 120 seconds, respectively. Exemplary results are shown in Figure 2. It should be noted that the ASM algorithm was performed on gradient images of CT data in our experiments.

VII. Summary

In the paper we presented the method of automatic landmarks propagation and the new algorithm of segmentation of prostate image in the CT scan exploiting the AAM technique. Efficiency of the proposed solution has been confirmed by visual examinations and compared with outlines performed by a radiotherapy specialist. In our future research we plan to use the presented approach to create a system for automatic segmentation using models of other organs like bladder, rectum pelvis etc.

ACKNOWLEDGMENTS

The research was supported by Polish grant founded by Polish Ministry of Science and Higher Education/National Science Centre, project NN 518 497739.

REFERENCES

[1] S. Chen, M. Lovelock, and R.J. Radke, "Segmenting the prostate and rectum in CT imagery using anatomical constraints," Medical Image Analysis, 15(1), pp. 1-11, February 2011.

[2] A. Skalski, et al., "Computed Tomography - based radiotherapy planning on the example of prostate cancer: Application of Level-Set segmentation method guided by atlas-type knowledge," Conf. ISABEL'11, ACM Digital Library, ISBN 978-1-4503-0913-4/11/10, 2011.

[3] S. Klein, et al., "Segmentation of the prostate in MR images by Atlas Matching," In 4th IEEE Int. Symposium on Biomedical Imaging From Nano to Macro, pp. 1300-1303, 2007.

[4] J. Dowling, et al., "Automatic MRI Atlas-Based External Beam Radiation Therapy Treatment Planning for Prostate Cancer," LNCS 6367, pp. 25-33, 2010.

[5] T,F, Cootes et al., "A trainable method of parametric shape description," Image Vision Comput. 10(5), pp. 289-294, 1992.

[6] T.F. Cootes, and C.J. Taylor, "Statistical models of appearance for computer vision," Technical Report, University of Manchester, 2004.

[7] X. Gao, et al., "A review of Active Appearance Models," IEEE Trans. on Systems, Man, and Cybernetics - Part C: Applications and reviews. 40 (2), pp. 145-158, 2010.

[8] T. Heimann, and H.P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," Medical Image Analysis, 13, pp. 543-563, 2009.

[9] A.F. Frangi, et al., "Automatic ASM Construction via Atlas-Based Landmarking and Volumetric Elastic Registration," LNCS 2082, pp. 78-91, 2001.

[10] C. Tobon-Gomez, et al., "Automatic Construction of 3D-ASM Intensity Models by Simulating Image Acquisition: Application to Myocardial Gated SPECT Studies." IEEE Trans. on Medical Imaging, 27(11), pp. 1655-1667, 2008.

[11] A. Skalski, A. Kos and T. Zielinski, "Using ASM in CT data segmentation for prostate radiotherapy", Int. Conf. on Computer Vision and Graphics, LNCS, Warsaw, September 24-26, 2012, accepted.

[12] S.K. Balci, P. Golland, and W.M. Wells, "Non-rigid Groupwise Registration using B-Spline Deformation Model," The Insight Journal, 2007, DOI-http://hdl.handle.net/1926/568

[13] A.F. Frangi, et al., "Automatic ASM Construction via Atlas-Based Landmarking and Volumetric Elastic Registration," LNCS 2082, pp. 78-91, 2001.

[14] A.F. Frangi, et al., "Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modeling'," IEEE Trans. on Medical Imaging, 21(9), pp. 1151-1166, 2002.

[15] S. Ordas, et al., "A statistical shape model of the whole heart and its application to model-based segmentation," SPIE Medical Imaging: Physiology, Function, and Structure from Medical Images, SPIE, . 6511, 2007.

[16] H.C. van Assen, et al., "SPASM: a 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data," Medical Image Analysis, 10(2), pp. 286-303, 2006.

[17] Y. Jeong, and R. Radke, "Reslicing axially sampled 3D shapes using elliptic Fourier descriptors," Medical Image Analysis, 11(2), pp. 197-206, 2007.

[18] T.F. Cootes, D. Cooper, C.J. Taylor, and J. Graham,: "Active Shape Models Their Training and Application," Computer Vision and Image Understanding. 61(1), pp. 38-59, January 1995.

[19] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," IEEE Trans. on Pattern Analysis and Machine Intelligence 23(6), pp. 681-685, June 2001.

[20] Rueckert, D., Sonoda, L.I., Hayes, C., et al., "Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images," IEEE Trans on Med. Imag., 18(8), pp. 712-721, 1999.

[21] G. Edwards, A. Lanitis, C. Taylor, and T. Cootes, "Statistical Models of Face Images Improving Specificity," Image and Vision Computing, vol. 16, pp. 203-211, 1998.

[22] D.J. Kroon: Active Shape Model and Active Appearance Model http://www.mathworks.com/matlabcentral/fileexchange/26706-active-shape-model-asm-and-active-appearance-model-aam

Noninvasive Articulograph. OMAP Architecture in Visual Signal Processing

mgr inż. Mirosław Sobotka
Faculty of Electronics and Information Technology
Warsaw University of Technology
Warsaw, Poland

prof. nzw. dr hab. inż. Antoni Grzanka
Faculty of Electronics and Information Technology
Warsaw University of Technology
Warsaw, Poland

Abstract — **The concept of USG based noninvasive articulograph has been presented for speech production registration. The proposed articulograph is a portable device with OMAP board as a signal processor unit. Combining the information about tongue position and labiograms the progress in linguistic theory of speech can be expected.**

I. Introduction

What is Articulography? There is no one definition of the therm of Articulography but shortly and simplifying we can define it as a measurement of moving parts (articulators) of speech apparatus in speech production process. The main problem was how to record it without interrupting the natural process of speech. The second problem is how to record movement of articulators located inside the speech apparatus. Logopedics heavily analyze the movement of lips. When it comes to analysis of tongue movement the procedure is much more complicated. The science tried to find the solutions for the problem by introducing palatography [1]. The method is very inconvenient for the patient. More convenient way was articulography introduced by German company called Carstens Medizinelektronik. The method is reasonable convenient for patient but the procedure is very complicated.

Working previously on subject of speech reconstruction for people with partial or full laryngectomy the idea of reconstructing of the speech from information other than audio data. Series of experiments let to conclusion that visual information coming from mouth is insufficient. The natural way is to extend visual channel by adding the data from tongue position. We should emphasis here that audio information was unavailable and whole procedure should be based on any other information channel. Experiments performed with USG devices brought promising results. Device based on USG and dedicated for tongue tracking does not exist on the market. This initiative has the goal of creating system for research purpose and maybe commercial usage as well. The interesting aspect is to verify if lack of positive back propagation of process influences the articulatory movements. The goal of this work is to find cheap and convenient for patient methodology of registration hidden articulators.

Inspiration for the project was palm sized USG devices offered by companies from China. Usage of mixed DSP and ARM architecture allows minimize the size of device and power consumption together with maximization of processing power.

The project is in very initial stage due to lack of the cheap USG head with well defined and standardized communication interface.

II. Scientific Goals

The main scientific goal is to verify ability of ultrasound to track the movement of the tongue. Other goal is to analyze the movements of articulators like tongue shapes and position in vocal tract during speech production. There is a great interest in evaluation of the tract geometry in speech research applications as well as in medicine and other application areas. In the early studies the X-ray radiography has been utilized, however this method is limited to small counts of subjects as the test is invasive and demands large equipment. The modern MRI, electromagnetic articulography, palatography pose similar disadvantages. Looking for the noninvasive, portable technique we propose ultrasonography as the method of tongue imaging. We will present the proposal of research combining of labiograms and USG for construction a portable device. There are few expected applications of the proposed instrument. The speech therapists would evaluate the quality of expression looking at the source. W are going to reconstruct the speech of persons after laryngectomy where USG is supposed to bring the key information. And finally we would like to evaluate the precision of the technique.

Once proven and implemented technology mentioned above the authors will focus on attempts to take dialog with theories like Distinctive Regions Model (DRM) [Mrayati 1988] and quantal theory of speech [Stevens 1989]

III. Meaning for Science and Industry

In speech production process there are some area of knowledge, especially in medical context, which require deeper knowledge. Main instrument for information coding in human voice communication is articulatory apparatus. This project will focus on vocal tract, articulators movements and voice – all participants in speech production. Most result of research on oral cavity during speech comes from invasive methods like palatography, electromagnetic articulography or radiography. Methods proposed in projects, based on ultrasonography are in initial stage in this area of usage [Stone 2005]. The technical part of this project will be a challenge and should produce significant volume of information as it will be evaluated on large set of patients. There is assumption, in speech communication the most significant role plays so called transient states. We can find the quantity of scientific activities, which were focused on mentioned above site of speech production process. All of them have deductional character and require more empirical verification. As sample of deductional model we can present Distinctive Regions Model DRM [Mrayati 1988]. DRM model is in the center of wider deductional approach leading to explanation of voice phenomenon including speech [Carré R 2009]. Example of success in building phonetic model based on deduction is publication [Liljencrants 1972], where author tries to explain phonetic structure of vowel system based on articulators model and perceptual contrast rule. The result could be achieved by the introduction of numerical interpolation of maximum contrast rule. Perceptual contrast has a long tradition in linguistic [Jakobson R 1941].

It is also worth to take challenge of research on quantal theory of speech [Stevens 1989]. By invoking acoustic stability criteria introduced vowel formants. This kind of research try to take assumption from the higher level, in kind of acoustic contrast maximization, minimum effort or simplicity. Speech production system and it's perception is treated as an effect of evaluational process. This process is driven by the main evaluational rules applied to human articulatory apparatus including vocal tract. Initially it was a tube adopted and formed later for communication process. Effort minimalism in control of communication was the major criteria which nature used to drive evaluational process of adaptation articulatory system. To bring something new in that knowledge, it will be defined new noninvasive method of registration of articulation. For many years such role played introduction of palatography. Palatography has proven, that thru the control of contact of tongue with palate, it is possible to estimate dysfunctions of speech and judge advances in removing them.

Some part of research activities will be focused on speech synthesis ability based on articulators state. Hidden Markov Model will be used for mentioned above task. Pioneers in this area using methods of electromagnetic articulography were scientific institution in Japan [Ling 2010]. In system where HMM computation is used to estimate articulators movements, the phase of learning is based on pre-recorded features of articulators linguistic context labeling. When text is complemented with acoustic features, the phase of HMM learning is started. The model is able to detect the dependencies acoustic and articulators movements. In speech synthesis process optimal trajectories of articulators are generated from learned models. Maksimum likelihood (ML) is used for computation of optimal trajectory [Tokuda 2000].

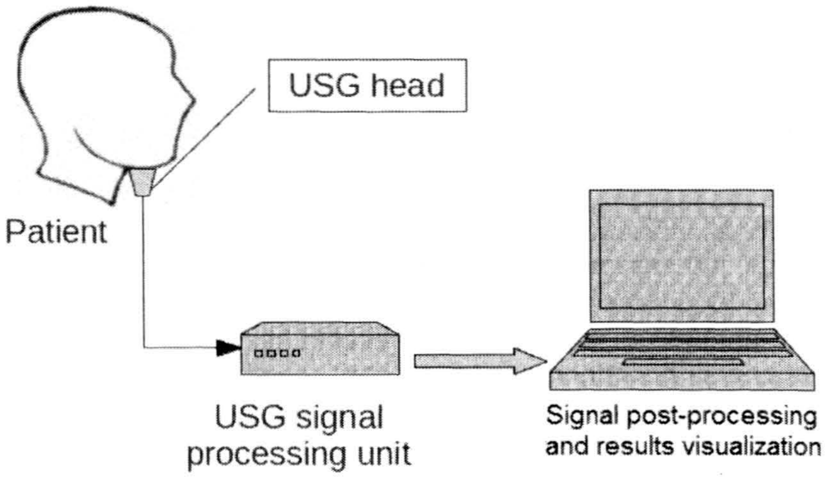
Other aspect of research is the attempt to build the system for reconstruction of the speech in situation when one element of vocal tract does not function properly. As example will be taken the case of patients after partial or full laryngectomy. These people use so called pseudo whisper. The approach is to reconstruct the voice from articulator movements

Building of the noninvasive system for analysis of articulatory movement will give researchers new device for verification of theories which have so far deductive character. Authors have also the ideas of usage modified version of the device in commercial products. Mainly it will target building equipment for reconstruction of the speech for people with laryngectomy.

IV. The Solution

A. Hardware and software

USG head should be placed under the jaw of patient. The mounting should be stable and convenient for patient. USG signal is interpreted by signal processing unit. Coordinates of tongue surface could be passed to dedicated PC or interpreted and visualized by OMAP board. Whole process can be synchronized with sound and video of lips movements. USG head is the simple linear type, ideally with USB interface. Target system should use USG head with two orthogonal lines of sensors. Cross like shape allows to record 2D position of the predefined points on the tongue. Third dimension is take based on the distance of the point from the USG head


USG head
Patient
USG signal processing unit
Signal post-processing and results visualization

In experiments following hardware and software was used:

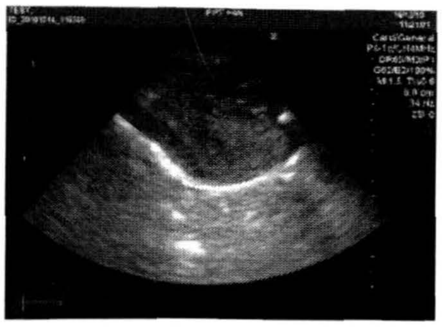
1. BeagleBoard (OMAP 3.5) from Texas Instruments. OMAP architecture offers hybrid architecture with multipurpose ARM processing unit and dedicated DSP (C64x family). Heavy algorithms can be optimized for DSP architecture. General purpose tasks (visualization) can implemented on ARM core. Next generation OMAP 4 architecture offers two general purpose ARM cores and better DSP unit. Evaluation board worked under Linux Angstrom control.

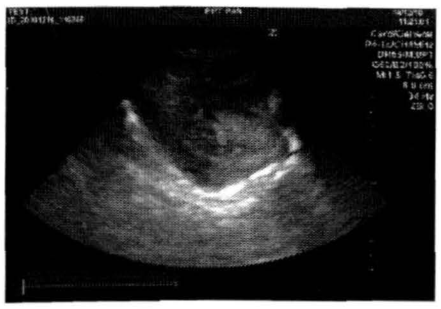
2.   USG device with USB interface called Sontrance (http://www.sontrace.com). It was provided by company called Draminski . It is simple and relatively cheap device but biggest advantage is USB interface. USB interface allows easy integration with OMAP evaluation board. Initial recording of tongue position were performed using standard hospital USG device.

3. The software implementation is done in C++. For DSP utilization, DSPLink from Texas Instruments was used. Compilation for the target system was done using dedicated for BeagleBoard Toolchains. As compiler gcc together Eclipse IDE was used.

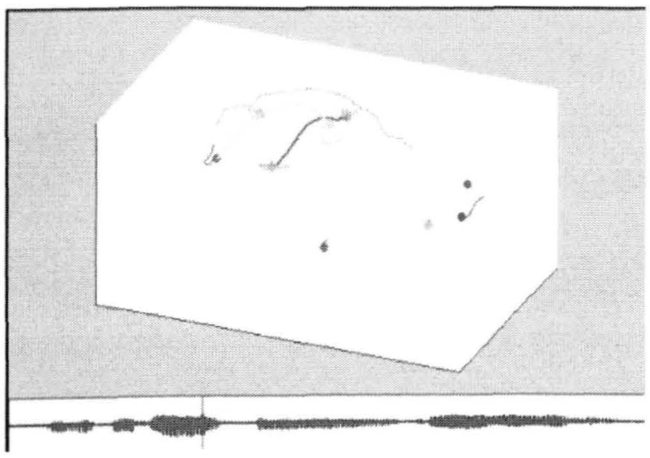
B. Results and final goals

Experiments were focused on the visualization of tongue surface. The surface of tongue has high impedance for ultrasound. As a consequence in USG images tongue is very well shown. Below is the example sequence of images for D vowel.





As the result of signal processing, the system should create the 4D (3D in time) representation of the articulatory movements



V. Applications

A. *Measurement tool*

The device should be able to produce result with accuracy to be sufficient for measurement purposes. The biggest problem is to define the coordinates system. The other obstacle is to guarantee the stability of coordinates system. As reference it can be used system AG500 German company called Carstens Medizinelektronik. (http://www.articulograph.de/)

Verification of the results should be performed by Logopedics and Phoniatrics institutions

B. *Device for reconstruction of the speech*

Thesis: Information about the shape of the lips and the tongue location should be sufficient to recognize the vowel. The sequence should lead to recognition of the word.

In this solution, the simplified version of the articulograph is used as system extending the support vector for classification of the vowel based on visual information from shape of the lips. The solution uses camera recording the lips movement and articulograph for the information about tongue position. The processing unit can be based on OMAP4 or combination of dedicated DSP system connected to PC. At the moment experiments are performed using OpenCV library optimized for TI TMS320 architecture and Support Vector Machine for classification.

VI. Why OMAP

OMAP architecture becomes recently more popular mostly due to development of smart phones. The major advantages of OMAP are:

1. The mixed architecture – general purpose ARM unit and specialized DSP unit

2. Relatively low power consumption. For the experiments, the author used BeagleBoard xM (OMAP 3.5) and PandaBoard (based on OMAP 4) evaluation boards. The power consumption of both boards varies from 5 to 10W. It is very satisfying result for fully autonomic system.

3. Strong support from open source community. OMAP based evaluation boards were able to run Linux distributions. It is also possible to find off-the-shelf specialized professional libraries like OpenCV.

4. OMAP based system can be build as fully autonomic system focused on particular processing tasks. Only results can be exchanged with master system.

5. Strong support from companies using OMAP in own System On Chip solutions. Very good example is Texas Instruments. The company offers professional libraries for DSP modules of own chips. Library like DSPLink can be obtained free of charge.

6. Low heat emission. Heat emission can be critical in medical application.

References

[1] Carré R: Carré R: Dynamic properties of an acoustic tube: Prediction of vowel systems. Speech Communication 51 (2009) 26–41

[2] Jakobson R: Kindersprache, aphasie und allgemeine lautgesetze. Uppsala. (1941), przedruk w Selected writings I, The Hague: Mouton (1962) 328-401.

[3] Liljencrants J, Lindblom B: Numerical simulation of vowel quality systems: the role of perceptual contrast. Language 48 (1972) 839–862.

[4] Mrayati M, Carre R, Guerin B: Distinctive regions and modes: a new theory of speech production. Speech Communication 7 (3), 1988: 257-286.

[5] Stevens, K.N: On the quantal nature of speech. J. Phonetics 17, (1989) 3–45.

[6] Stone M: A guide to analysing tongue motion from ultrasound images. Clinical Linguistics and Phonetics 19, 2005: 455–502

[7] Ling Z.H, Richmond K, Yamagishi J: An Analysis of HMM-based prediction of articulatory movements. Speech Communication 52, 2010: 834–846

[8] Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T: Speech parameter generation algorithms for HMM-based speech synthesis. In: ICASSP, 2000, vol. 3, pp. 1315–1318

[9] Speech Communication 2009, Thomas Hueber, Elie-Laurent Benaroya, Ge ´rard Chollet, Bruce Denby, Ge ´rard Dreyfus, Maureen Stone, Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips

[10] Tim Bressmann, Quantitative Assessment of Tongue Shape and Movement Using Ultrasound Imaging

[11] Mirko Grimaldi, Barbara Gili Fivela, Francesco Sigona, Michele Tavella, Paul Fitzpatrick, Laila Craighero, Luciano Fadiga, Giulio Sandini, Giorgio Metta, New Technologies for Simultaneous Acquisition of Speech Articulatory Data: 3D Articulograph, Ultrasound and Electroglottograph

[12] Lisa Tang, Ghassan Hamarneh, Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization

[13] Lisa Davidson 2006, Comparing tongue shapes from ultrasound imaging using smoothing splin eanalysis of variance

[14] Michael Aron, Marie-Odile Berger,Erwan Kerrien 2008, Multimodal Fusion of Electromagnetic, Ultrasound and MRI Data for Building an Articulatory Model

[15] Mathews Jacob, Heike Lehnert-LeHouillier, Sourabh Bora, Stephen McAleavey, Diane Dalecki, and Joyce McDonoug 2008, Speckle Tracking for the Recovery of Displacement and Velocity Information from Sequences of Ultrasound Images of the Tongue

[16] Interspeech 2008, Chao Qin, Miguel A.Carreira-Perpianan, Korin Richmond, Alan Wrench, Steve Renals, Predicting tongue shapes from a few landmark locations

[17] Interspeech 2007, Thomas Hueber, Gérard Chollet, Bruce Denby, Gérard Dreyfus, Maureen Stone, Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips

[18] Ian Fasel & Jeff Berry, Deep Belief Networks for Real-Time Extraction of Tongue Contours from Ultrasound During Speech

[19] Maureen Stone 2004, A Guide to Analysing Tongue Motion from Ultrasound Images

[20] Alan A. Wrench and James M. Scobbie, High-speed Cineloop Ultrasound vs. Video Ultrasound Tongue Imaging: Comparison of Front and Back Lingual Gesture Location and Relative Timing

[21] Michael Aron, Nicolas Ferveur, Erwan Kerrien, Marie-Odile Berger, Yves Lapri, Acquisition and synchronization of multimodal articulatory data

SESSION 2:

AUDIO PROCESSING I

Automatic Analysis System of TV Commercial Emission Level

Paweł Spaleniak, Bożena Kostek
Multimedia Systems Department
Gdansk University of Technology
Gdańsk, Poland
{papol, bozenka}@sound.eti.pg.gda.pl

Abstract — **The purpose of the study was to determine whether the level of commercial emission is higher than the level of regular program and to check if the commercials broadcasters follow the recommended levels of loudness. The paper shortly reviews some chosen methods of volume measurements specified in the ITU and EBU recommendations. Then, it describes a prototype of a system implemented in Embarcadero C++ Builder 2010 which carries out automatic evaluation of loudness using the recordings acquired directly from TV programs. In the end, the results of the measurements obtained for TV commercials are shown, and the conclusions are drawn. The final Section outlines also future work being planned.**

Keywords — ***Content classification, Detection of TV commercials, Loudness measurements***

I. Introduction

The problem of too high level of TV commercial emission level can be observed in almost all countries. Watching programs interrupted by loud commercials, adverts is burdensome and significantly affect the viewer's comfort. This problem relates not only to sound but also to the image. Commercials are very often broadcasted with a higher contrast and with more frequently shot changes. Those actions have to attract the attention of a potential client. In Poland, the KRRiT (*National Broadcasting Council*) regulation on program emission level [6] sets that the emission level should be determined on the basis of the multichannel sound level objective measurement algorithm. The algorithm proposed by KRRiT is based on the ITU (*International Telecommunication Union*), BS.1770-1 recommendation [8].

This paper includes a short review of some methods and recommendations on analyzing emission level of audio-video signals and describes the developed Automatic Analysis System of the TV Commercial Emission Level.

II. Parameters Definitions

Before a brief description of chosen loudness measurement methods is given, some of the parameters should first be recalled:

- emission level - sound pressure level, expressed in LUFS or LKFS, used to describe loudness of TV programs;

- loudness - attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud [1];

- sound pressure level - logarithmic measure of the effective sound pressure of a sound relative to a reference value;

- dB SPL - decibel unit used to describe sound pressure level;

- LKFS - Loudness, K-weighted, relative to full scale - loudness standard designed to enable normalization of audio levels for delivery of broadcast TV and other video [4]. LKFS is standardized in ITU-R BS.1770 [8];

- LUFS - EBU loudness unit, related to full scale, described in EBU R 128 [3].

III. Description of Chosen Loudness Measurement Methods

A. Methods and regulations on loudness measurement

This Section describes some of the methods concerning the TV program loudness measurement. First, it should be noted that all official recommendations use loudness parameter in contradiction with its physical definition. However, to maintain consistency, the authors of this paper use loudness parameter in the same context as recalled in the ITU-R.1770-1 recommendation. The ITU-R.1770-1 and the EBU R-128 recommendations are the basis for the objective estimation of loudness and provides requirements for the loudness measurement method.

B. Equivalent Sound Level Measurement

The equivalent sound level L_{Eq} is defined as the level of the steady state audio signal that has, over a given period, the same energy as a fluctuating sound. The equivalent sound level is commonly used in measuring environmental noise pollution but it can also be employed to measure loudness (using *Revised Low Frequency B*-weighting curve; RLB) of mono audio signals [8]. The revised RLB analyzer implements a method of sound level measurement especially designed for the monitoring of broadcast audio level. RLB is similar to B-weighting, except that the high frequency transfer function is flat instead of having a high frequency roll-off. This spectral weighting was found to be an effective objective measurement of loudness of typical broadcast material by Soulodre (2004) [10], and is recommended for use by the International Telecommunication Union [2][8]. It is a fully functional method, which requires low computing resources and gives satisfying results. L_{Eq} is defined as follows:

$$L_{Eq}(W)=10log[\frac{1}{T}\int_0^T \frac{x_W^2}{x_{Ref}^2}dt] \quad (1)$$

where: W - used weighted curve,
x_W - signal at the filter input,
x_{Ref} - reference level.

The block diagram of the equivalent sound level (*RLB*) meter is shown in Figure 1.

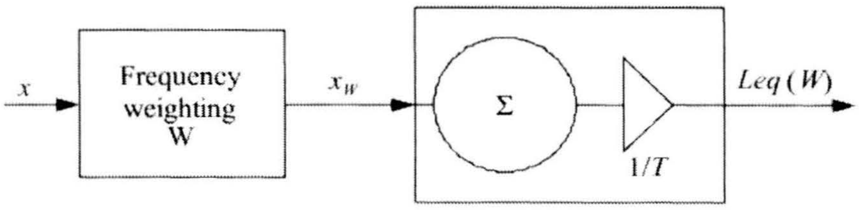


Figure 1. Block diagram of equivalent sound level (*RLB*) meter

C. Recommendation ITU-R, BS.1770-1

International recommendation ITU-R, BS.1770-1 [8] describes audio signal (mono and multichannel) measurement algorithms, which are used to present subjective loudness in numerical form. The measurement procedure does not include the low frequency effect (LFE) channel.

Figure 2 shows a block diagram of loudness measurement in multichannel system. The first step of the algorithm developed by ITU is pre- filtering of each channel. The pre-filter characteristics is shown in Figure 3, the block diagram of this filter is presented in Figure 4. In Fig. 3 the first stage of the pre-filtering takes into account the acoustic effects of the head (high frequency acoustic waves reflections), where the head is modeled as a rigid sphere [8].

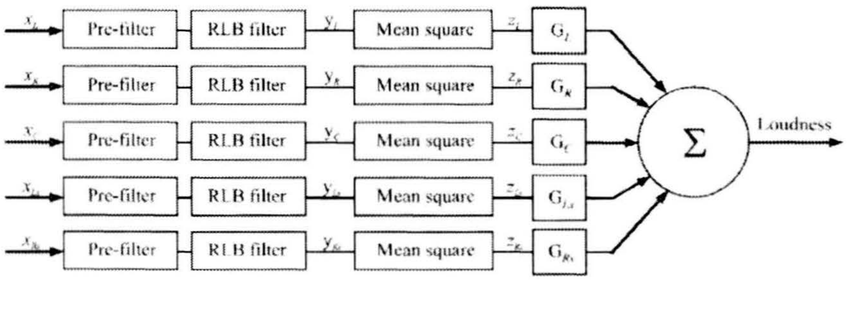


Figure 2. Block diagram of the loudness measurement according to ITU, BS.1770-1 [8]

Table 1 contains pre-filter coefficients, appropriate only for the 48kHz sampling frequency. In other cases, different coefficient values should be chosen to obtain characteristics consistent with Figures 3 and 4.

TABLE I. TABLE OF PRE-FILTER COEFFICIENTS TO MODEL THE ACOUSTIC EFFECTS OF THE SPHERICAL HEAD

-	-	b_0	1.5351248595697
a_1	-1.69065929318241	b_1	-2.69169618940638
a_2	0.73248077421585	b_2	1.19839281085285

The second step of the algorithm is high pass filtration which uses the RLB weighting curve. Figure 5 shows RLB weighting curve. The block diagram is the same as initial filter (Fig. 4). Filter coefficients are presented in Table II.

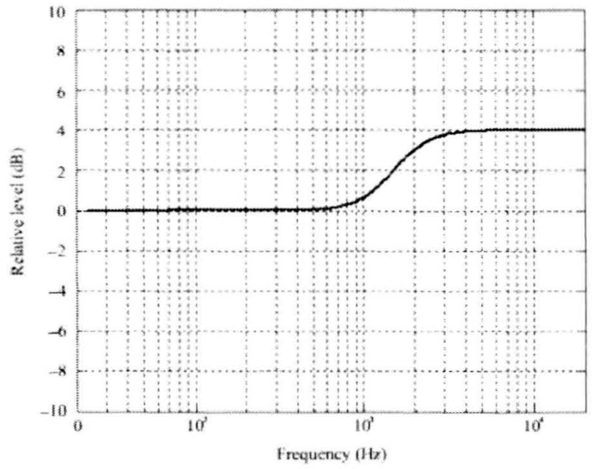


Figure 3. Pre-filter frequency characteristics taking the acoustic effects of the head influence into account [8]

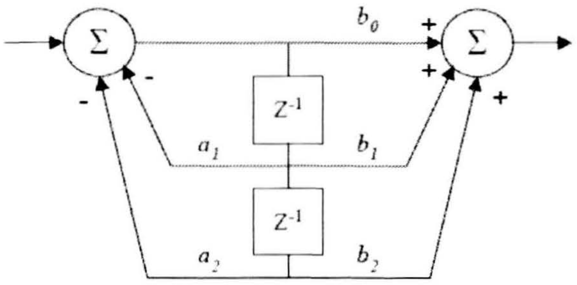


Figure 4. Block diagram of pre- and RLB filters of 2nd order [8]

TABLE II. TABLE OF THE RLB FILTER COEFFICIENTS

-	-	b_0	1.0
a_1	-1.99004745483398	b_1	-2.0
a_2	0.99007225036621	b_2	1.0

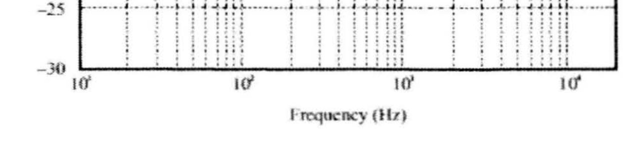


Figure 5. RLB filter frequency characteristics

The next step of the process is the calculation of the root mean square value of energy, in time period *T* (2). The optimum time interval *T* is not proposed in the recommendation.

$$z_i = \frac{1}{T}\int_0^T y_i^2 dt \quad (2)$$

where: y_i - signal after both filtration operations,
i - *L, R, C, Ls, Rs* (all used channels).

After calculating z_i for all of the channels the loudness can be computed according to (3). The loudness unit is LKFS or LU (1LKFS = 1LU = 1dB) [8].

$$loudness = -0.691 + 10 \log \sum_i^N G_i z_i \qquad (3)$$

where: N - number of active channels
G_i - channel weighting coefficient (helps to account source location relative to the head)

D. EBU R-128 recommendation

The algorithm proposed by European Broadcasting Union [3] organization expands the ITU algorithm by adding some new parameters, for example: Program Loudness, Loudness Range or True Peak Level. The loudness measure process is similar to the ITU, BS.1770-1 algorithm, however the gating block is added. However, it should be added that recently, the revised version of the ITU recommendation (BS.1770-2) has been prepared in which the gating block was applied [9]. The gating function prevents taking low-level signals into account. EBU recommendation also changes the loudness unit from LKFS to LUFS.

IV. Assumptions of the Designed System

The main purpose of the system is to see whether commercials are louder than other programs. The system assumptions are as follows:

- Windows application,
- automatic detection of commercials in an analyzed signal and information the user about the beginning and the end of a commercial block,
- implementation of the ITU and EBU loudness measure algorithms according to recommendations,
- an offline mode implementation, which enables the user to import an audio signal, extracted from TV recording (stereo *wave* format, 16bit, 48kHz),
- Presentation of the loudness to the user after the measurement process is completed.

V. Description of the Designed System

The automatic Analysis System of TV Commercials Emission Level is, according to assumptions, a Windows application. It has implemented both, ITU and EBU algorithms. In the current version the system does not have the online mode. The application engineered detects commercials in the imported audio track and measures loudness and loudness range (for the EBU algorithm).

The analysis starts with selecting an algorithm. The next step is the audio file selection. After the file is selected, the analysis process starts. Figures 6 and 7 show the application interface during the analysis process. All of the results are saved in a single text file, which enables the user to find out if the commercial block is emitted at a higher level than the other TV program content. The used detection algorithm is based on music recognition algorithm developed by the Shazam Entertainment Ltd [12]. The method uses only audio tracks for analysis, which enables the developer to implement it in both, TV and radio commercial detection.

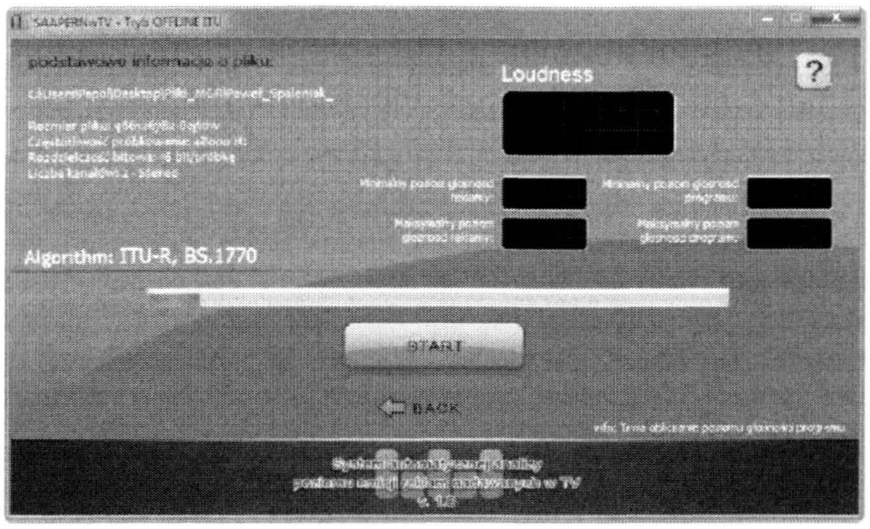
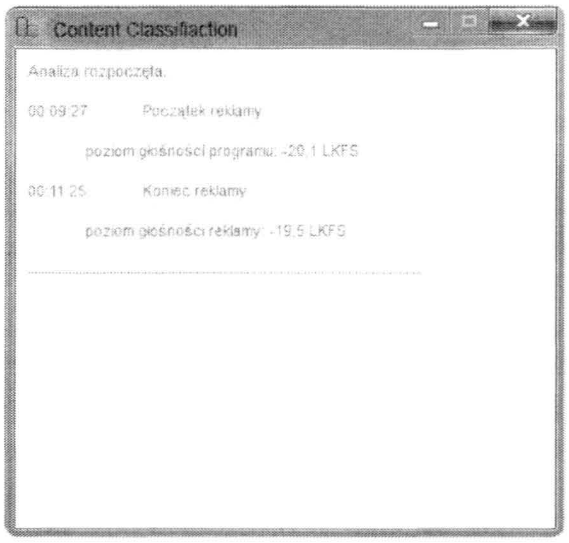


Figure 6. The main application window during the analysis process



Analiza rozpoczęta.

00:09:27 Początek reklamy

poziom głośności programu: -20,1 LKFS

00:11:25 Koniec reklamy

poziom głośności reklamy: -19,5 LKFS

Figure 7. An additional window showing information about beginning and the end of commercial

The basis of implemented algorithm is the *spectral fingerprint* of a commercial jingle. *Fingerprint* description must be created manually for each TV channel. The idea of the *spectral fingerprints* is based on time offsets between characteristic points of the spectrogram. Figure 8 and Table III show the principal ideas behind the system working.

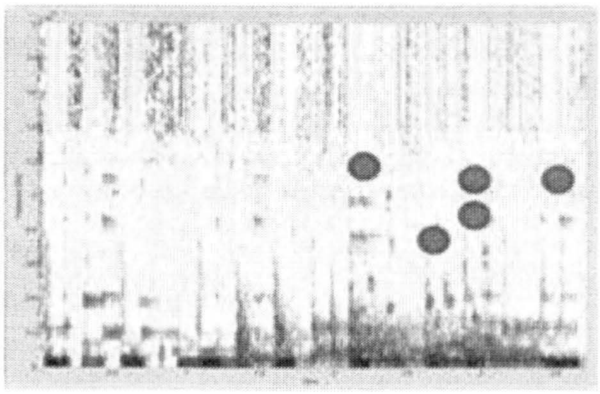


Figure 8. A commercial jingle spectrogram with marked characteristic points

TABLE III. FREQUENCIES AND TIME OFFSETS OF EXAMPLE FINGERPRINT CHARACTERISTIC POINTS

Frequency [Hz]	Time offset [sample]
1250 (anchor point)	0

1250	494
1250	1614
1250	2834
2500	2844
3750	4254

Implemented algorithm is searching for the anchor point of the prepared *fingerprint*. Once the anchor point is detected, the application starts to "listen" for the rest of the characteristic points. If all of the points are detected, the system shows information about beginning (or about the end) of a commercial block. The fewer number of points, the faster analysis.

The application uses 512-point FFT (window type: rectangular; window length: 1024 samples - without overlapping) to create the audio spectrogram. In this configuration the algorithm gives satisfactory results - about 90% effectiveness in detection of commercials. It is obvious, that it also depends on precision of the *fingerprint* description. The effectiveness can be increased by using other than rectangular analysis windows, for example Hamming window.

The application returns loudness of a whole analyzed file, loudness of each commercial block and program, and in addition for the EBU algorithm: loudness range.

VI. Experiments

In all of the experiments audio tracks extracted from the TV recordings collected in database were used. The database contains recordings from four Polish TV broadcasting channels. The material was recorded using the stationary DVD recorder (Frame size: 720x576 pixels; Audio: 48kHz, 16bit, 2 channels) and standard cable TV signal for a source. The storage format is DivX coded *avi* with the same frame and audio parameters. The total length of stored material is14 hours, 57 minutes and 30 seconds.

A. Algorithm comparison - loudness with ITU and EBU

The first test was a simple comparison of the two measurement algorithms implemented. Randomly selected fragments of TV recordings were used to compare loudness measurement results by using both methods. Table IV shows some of the results.

TABLE IV. Loudness measurements with ITU and EBU algorithms (top broadcaster)

File name	Loudness	
	ITU [LKFS]	*EBU* [LUFS]
01 12 2010_09 40 c3	**-19.6**	**-18.9**
01 12 2010 09 40 c4	**-19.9**	**-19.5**
01_12 2010 09 40 c5	**-19.4**	**-19**
01 12 2010 09 40 c6	**-19.1**	**-18.7**
01 12 2010 09 40 c7	**-19.5**	**-19.3**
01 12 2010 09_40 c8	**-19.6**	**-19.3**

Test results were consistent with expectations. In all cases loudness was higher for the EBU algorithm. It is caused by the implemented gating block. This block causes the measurement to stop when the level of input signal is lower than the threshold properly set. However, an interesting result is the fact that all levels are higher than acceptable -23 LUFS.

B. Comparison of commercial and program loudness

The aim of this test was to find out if the commercials are emitted at a higher level than actual program. The algorithm used in this test is based on the EBU R 128 recommendation. Table V shows some of the results obtained.

TABLE V. Loudness measurements with ITU and EBU algorithms (top broadcaster)

	Commercial	**Program**
	1	
Loudness:	-18.5	-22.2
LRA:	6.2	13.1
	2	
Loudness:	-19	-18.7
LRA:	4.8	5.8
	3	
Loudness:	-18.3	-20
LRA:	5.9	13.5
	4	
Loudness:	-18.5	-20.8
LRA:	4.8	6.4
	5	
Loudness:	-18.4	-20
LRA:	5	7
	6	
Loudness:	-19	-20.6
LRA:	5.4	6.3

The results were similar in most cases. Commercials were emitted at a higher level than the programs indeed. Analyzing the loudness range parameter (LRA) differences in the results obtained can be seen. It is dependent on the content of recording. Narrow ranges of LRA (5-6 LU) characterize talk-show type programs (no music or special effects); medium ranges (7-10 LU) - Polish series; wide ranges (>10 LU) - action movies.

Neither the programs nor the commercials do meet the requirements related to loudness (-23LUFS). This problem can be recognized in all of the broadcast stations material [11]. However, this may be caused to some degree by conversion of TV signal before it reaches the DVD recorder input.

VII. SUMMARY AND CONCLUSIONS

The main purpose of the application designed was to show loudness differences between TV commercials and programs. It enables the user to check if the commercial is emitted at a higher level compared to the adjacent regular program. Two approaches were employed based on ITU and EBU recommendations.

The processes of the TV signal transformations before it reaches DVD recorder are not known. To achieve the most reliable results the DVB-T signal in stream form should be used. However, the analyzed recording showed that commercial blocks are emitted at a higher level than the program content which is often found by other studies (e.g. [6]).

The Automatic Analysis System of TV Commercials Emission Level is at the stage of improving its functionality features. The focus is on automatic *spectral fingerprint* extraction; the future work will combine audio and video methods for detection process; real-time loudness analysis - these are the most important aspects. Since there are no commercially available applications that enable to automatically detect commercials in the TV signal and simultaneously measure loudness (in compliance with recommendations), thus this goal is worth pursuing.

References

[1] American National Standards Institute, "American national psychoacoustical terminology" S3.20, 1973, American Standards Association.

[2] D. Cabrera, S. Ferguson and E. Schubert, "'PsySound3: software for acoustical and psychoacoustical analysis of sound recordings,' Proceedings of the 13th International Conference on Auditory Display, Montreal Canada, June 26-29 2007, pp. 356-363.

[3] EBU TECH 3341, "Loudness Metering: 'EBU Mode' metering to supplement loudness normalization in accordance with EBU R 128", Geneva, Switzerland, December, 2010.

[4] R. van Everdingen, E. M. Grimm, M.J.L.C Schöpping, "Toward a Recommendation for a European Standard of Peak and LKFS Loudness Levels", Technical Paper, April, 2010.

http://www.grimmaudio.com/whitepapers/Toward%20a%20Recommendation%20of%20Peak%20and%20LKFS%20Levels%20-%20SMPTE%20journal.pdf

[5] M.C. Lavoie, S.G. Norcross, G.A. Soulodre, "The Subjective Loudness of Typical, Program Material", in Proc. of the AES 115th Convention, 2003.

[6] National Broadcasting Council, "Concerning Principles of Advertising and Teleshopping in Radio and Television Program Services", (Dz. U. 22 .07.2011 r.), PL, June, 2011.
(http://www.krrit.gov.pl/Data/Files/ public/pliki/regulations/3june2004 glosn a rekl.pdf)

[7] S. G. Norcross, M.C. Lavoie, "Loudness Normalization of Wide Dynamic Range Broadcast Material", 132 Audio Eng. Soc. Convention, Budapest, Paper No. 8606, 2012.

[8] Recommendation ITU-R BS.1770-1, "Algorithms to measure audio programme loudness and true-peak audio level", Geneva, Switzerland, April, 2006.

[9] Recommendation ITU-R BS.1770-2, "Algorithms to measure audio programme loudness and true-peak audio level", Broadcasting service (sound) 03/ 2011 (http://www.itu.int/dms pubrec/itu-r/rec/bs/R-REC-BS.1770-2-201103-I!!PDF-E.pdf)

[10] G.A. Soulodre and M. C. Lavoie (2006) "Development and evaluation of short-term loudness meters," 121st Audio Engineering Society Convention, San Francisco.

[11] P. Spaleniak, "Automatic Analysis System of TV Commercial Emission Level", M.Sc. thesis, Multimedia Systems Department, Faculty of Electronics, Telecommunications and Informatics, Gdańsk Univ. of Technology, Gdańsk, 2011.

[12] A.L. Wang, "An Industrial-Strength Audio Search Algorithm" (http://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf)

Application of Intrinsic Time-Scale Decomposition in Analyzing Sigma-Delta Modulator for Audio DAC

Marcin Lewandowski
Electroacoustic Division, Institute of Radioelectronics
Warsaw University of Technology
Warsaw, Poland
Marcin.Lewandowski@ire.pw.edu.pl

***Abstract* — The paper introduces a new approach for analyzing and processing non-stationary audio signals obtained from underlying nonlinear systems such as digital sigma-delta (ΣΔ) audio DACs. Their parameters and performance depend mainly on features of digital ΣΔ modulators. The Intrinsic Time-Scale Decomposition (ITD) method can self-adaptively decompose input signal into a monotonic trend (baseline signal) and a set of proper rotation components (PRCs) for which instantaneous parameters of signal are well defined. Finally, correlations between quantization noise and input signal (in particularly noise modulation) in digital ΣΔ modulators and possibility of the ITD method application for analyzing noise modulation is investigated.**

***Keywords* — digital ΣΔ audio DAC, digital ΣΔ modulator, noise modulation, intrinsic time-scale decomposition**

I. Introduction

Nowadays, ΣΔ audio analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) are commonly used in both consumer and professional audio equipment. Large number of psycho-acoustic tests showed that the quality of sound reproduced by ΣΔ audio DACs is worse than quality of sound reproduced by conventional PCM audio DACs. Preliminary analyses indicate that audio quality reduction of sound reproduced by ΣΔ audio DACs is strongly connected with noise modulation in digital ΣΔ modulator. Noise modulation is the effect where the noise floor of the modulator changes as the input signal changes i.e. quantization noise is correlated with input signal. According to [1-3] the variation in quantization noise should be less than 1 dB in order to be inaudible. Thus, for high quality audio applications the objective is to have a constant quantization noise that results in predictable audio quality. Therefore, noise modulation should be minimized or avoided if possible [4-10]. The main techniques that have been proposed to minimize or eliminate noise modulation include application of dither inside the digital ΣΔ modulator structure and selecting a digital ΣΔ modulator's feedback loop filter, which makes the modulator chaotic [11]. In this paper, only an application of dither signal inside ΣΔ modulator structure will be considered and quantization noise will be referred as error or quantization error.

Though the use of dither to prevent noise modulation in PCM systems is well-understood [12], there are many intriguing questions that remain when dither is applied to the digital ΣΔ modulator. That is mostly because of feedback loop in digital ΣΔ modulators that affects the probability distribution of the input to the quantizer in complicated ways [13]. The purpose of this paper is to look more closely at the noise modulation in digital ΣΔ modulators, introduce a new approach (called Intrinsic Time-Scale Decomposition) for analyzing and processing non-stationary signals obtained from underlying nonlinear systems and propose a new procedure for noise modulation analysis in digital ΣΔ modulators with ITD method applied.

II. Dither and Noise Modulation in Digital ΣΔ Modulators

Recent treatments of dither signal in digital ΣΔ modulators are that there has been wealth of analysis of such systems, with seemingly contradictory conclusions. Several authors have suggested that digital ΣΔ modulators may be self-dithered [14,15], but ability of self-dithering to minimize or eliminate noise modulation is unclear [16]. Alternatives to dithering have been proposed, including bit-flipping [17,18] and chaotic systems [19-23], but these results have yet to be put into the proper framework [24]. Noise modulation in digital ΣΔ modulators was also pointed out in [10,13].

As mentioned previously, the results of dither distributions and the resultant quantization error for PCM systems do not apply for digital ΣΔ modulators because the input to the quantizer includes the system input and noise shaped error. This statement was proven in paper [24] for simple digital ΣΔ modulator with one and multibit quantizer. Authors have shown how error moments depend on input to digital ΣΔ modulator.

In most digital ΣΔ modulators with 1-bit quantizer, addition of dither has no effect on the conditional moments of error. From Fig. 1, the input dependence in the error is clearly seen. These results actually represent six simulated systems (all with dc input signals): 1st and 2nd order digital ΣΔ modulators, each without dither, with rectangular dither at the quantizer, and with triangular dither at the quantizer. [24].

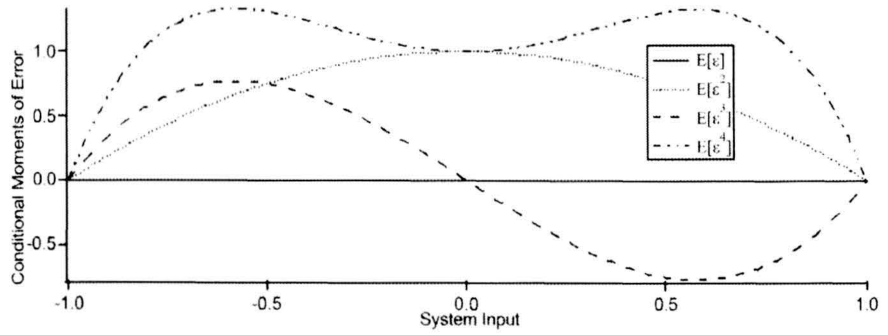


Figure 1. The first four conditional moments of error for digital ΣΔ modulator with a 1-bit quantizer [24].

In Fig. 2, the second and third order conditional moments of error are plotted as a function of dc input signal for digital ΣΔ modulators with multibit quantizer without dither, with dither applied to the input, and with dither applied to the quantizer. Although application of dither at the system input is rarely performed in practical applications, it is useful for illustration and serves to demonstrate why dither placed elsewhere in the feedback loop is often more beneficial [24]. The results show that application of rectangular dither in digital ΣΔ modulators with multibit quantizer successfully eliminates much of the input dependence in the error.

Another statistical analysis of error signal is presented in [25]. Authors examined dithered multibit 1st order digital ΣΔ modulator and considered all signals around the noise shaper. Numerous simulations including power spectral density (PSD), probability density function (PDF) and characteristic function (CF) calculation of signals were performed with different types and amplitude of dither signal. Authors showed that properly dithered multibit digital ΣΔ modulators can eliminate most of the noise modulation, but 1-bit digital ΣΔ modulators cannot be linearized by the use of dither. Input dependence on error was also described in [26] where testing for noise modulation was done by applying several dc levels as input signal to digital look-ahead based modulators, and for each dc level low-pass filtering the output and calculating the second order moment of the error. Alternatively, authors used SINAD calculation for varying sine wave's level as input signal to modulators. Simulations showed that both baseband noise power and SINAD varied with changes in input signal.

All presented simulations' results were estimated by averaging signals from digital ΣΔ modulators over time, thus even they prove the existence of noise modulation, short term behavior (from time to time) of digital ΣΔ modulator is averaged and clearly missing. Moreover, most of the simulations were conducted using dc level input signals, which are not used in practical audio applications. In the next two sections of this paper, a proposition of the new procedure for analyzing correlations between error and input signal in digital ΣΔ modulators with real world audio input signals will be introduced.

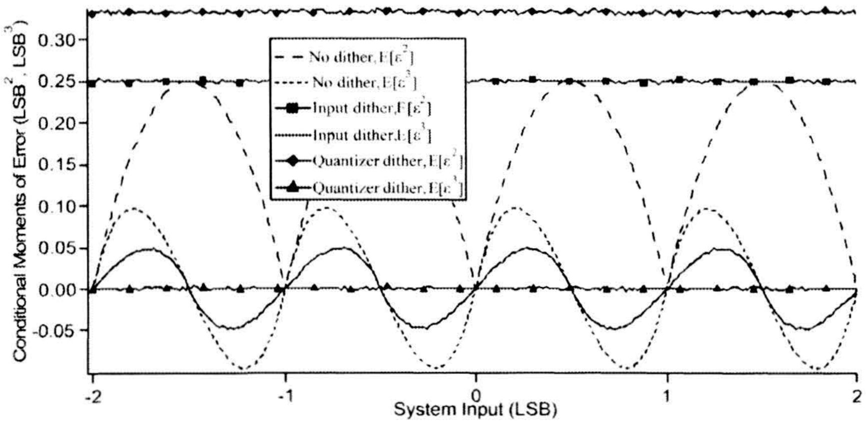


Figure 2. Second and third order moments of error as a function of input for a multibit first order digital ΣΔ modulator without dither, with rectangular dither at input, or rectangular dither before quantization [24].

III. THE PRINCIPLES OF ITD METHOD

The ITD method [27] can self-adaptively decompose input signal into a monotonic trend and a set of proper rotation components (PRCs) for which instantaneous parameters of signal are well defined. It constructs the piece-wise linear baseline signal between successive extrema and computes the instantaneous amplitude and frequency based on the piece-wise wave of each PRC.

Given a real-valued signal $\{X_t, t \geq 0\}$ an operator Λ is defined, which extracts a baseline signal from X_t in a manner that causes the residual to be a proper rotation (all minima are negative and all maxima are positive values). Signal X_t can be decomposed as

$$X_t = \Lambda \cdot X_t + (1-\Lambda) \cdot X_t = L_t + H_t, \qquad (1)$$

where $L_t = \Lambda \cdot X_t$ is the baseline signal and $H_t = (1-\Lambda) \cdot X_t$ is a proper rotation.

Let $\{\tau_k, k = 1,2,...\}$ denote the local extrema of X_t and for convenience define $\tau_0 = 0$. In the case of intervals on which X_t is constant, but which contain extrema due to neighbouring signal fluctuations, τ_k is chosen as the right endpoint of the interval.

Suppose that $X_k = X(\tau_k)$ and $L_k = L(\tau_k)$ have been defined on $t \in [0, \tau_k]$ and that X_t is known for $t \in [0, \tau_{k+2}]$. Then, baseline signal can be defined on the interval $(\tau_k, \tau_{k+1}]$ as follows:

$$\Lambda \cdot X_t = L_t = L_k + \left(\frac{L_{k+1} - L_k}{X_{k+1} - X_k} \right) \cdot \left(X_t - X_k \right), \quad t \in (\tau_k, \tau_{k+1}], (2)$$

where

$$L_{k+1} = \alpha \cdot \left[X_k + \left(\frac{\tau_{k+1} - \tau_k}{\tau_{k+2} - \tau_k} \right) \cdot \left(X_{k+2} - X_k \right) \right] + (1-\alpha) \cdot X_{k+1} \quad (3)$$

and $0 < \alpha < 1$ is typically fixed with $\alpha = 0.5$. Proper rotation component can also be defined as:

$$\mathrm{H} \cdot X_t \equiv (1 - \Lambda) \cdot X_t = H_t = X_t - L_t . \quad (4)$$

Once the input signal has been decomposed into baseline (L_t) and proper rotation component (H_t) as depicted in Fig. 3, with latter representing the highest relative frequency present in the input signal, the procedure can be re-applied using the baseline signals as input. This process can be repeated until a monotonic baseline signal is obtained. ITD decomposes the raw signal into a sequence of proper rotations of successively decreasing instantaneous frequency at each subsequent level of the decomposition:

$$X_t = \left(\mathrm{H} \cdot \sum_{k=0}^{p-1} \Lambda^k + \Lambda^p \right) \cdot X_t \quad (5)$$

Once the input signal has been decomposed, instantaneous amplitude, phase and frequency information can be extracted from proper rotation components. This instantaneous time-frequency-energy (TFE) information can be defined in a piece-wise manner, on each time-interval between successive up-crossings of a proper rotation, and based only upon information about the single wave of a proper rotation occurring during that period:

$$\theta_t = \begin{cases} \arcsin\left(\dfrac{x_t}{A_1} \right), & t \in [t_1, t_2), \\ \pi - \arcsin\left(\dfrac{x_t}{A_1} \right), & t \in [t_2, t_3), \\ \pi - \arcsin\left(\dfrac{x_t}{A_2} \right), & t \in [t_3, t_4), \\ 2\pi + \arcsin\left(\dfrac{x_t}{A_2} \right), & t \in [t_4, t_5), \end{cases} \quad (6)$$

where $A_1 > 0$ and $A_2 > 0$ are respective amplitudes of the positive and negative half-waves (portion of signal between adjacent zero-crossings) between successive zero up-crossings and $t_1 .. t_5$ are shown in Fig. 3. Instantaneous amplitude is piece-wise constant and determined by the extrema values of the proper rotation components between zero-crossings and instantaneous frequency are calculated as follows:

$$A_t = \begin{cases} A_1, & t \in [t_1, t_3), \\ A_2, & t \in [t_3, t_5), \end{cases} , \qquad f_t = \frac{1}{2\pi} \cdot \frac{d\theta_t}{dt} \quad (7)$$

where authors in [27] calculate f_t by differentiating instantaneous phase angles θ_t using an 11 coefficient least-squares FIR differentiating filter [28].

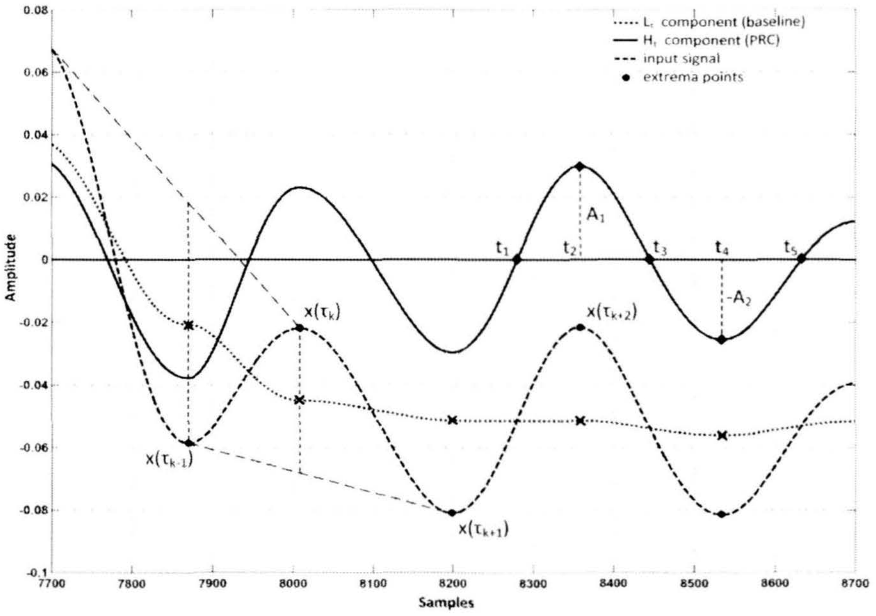


Figure 3. Illustration of ITD's extraction of the baseline and proper rotation component from an input signal.

The main characteristics that make the ITD method well suited for analysis of noise modulation in digital ΣΔ modulators are:

- precise temporal information regarding instantaneous frequency and amplitude of proper rotation component signals with a temporal resolution equal to the time-scale of occurrence of extrema in the input signal,
- ability to adapt to any time-scale and to use complete signal information, including all critical points such as inflection points and zero-crossings and not just local extrema in the input signal, thereby allowing weak signals embedded in stronger signals to be extracted.

IV. PROPOSED PROCEDURE FOR DIGITAL ΣΔ MODULATOR ANALYSIS

Proper analysis of digital ΣΔ modulator behavior with real world audio input signal would require very careful and precise analysis of input signal, all internal signals in digital ΣΔ modulator structure and output signal from the modulator. Moreover, those analyses should ensure that all short-term (temporal) changes in signals will be preserved. Previously introduced ITD method seems to be appropriate for such analysis. In Fig. 4 a model of 1st order digital ΣΔ modulator with one or multibit quantizer is presented along with blocks that refer to proposed analysis approach.

Input, B-bits audio signal $x(n)$ with $f_s = 44.1$ kHz is oversampled with $L = 64, 128, 256$ ratios, fed to the very precise low-pass filter and simultaneously to the digital ΣΔ modulator. The output signal $y(n)$ from digital ΣΔ modulator is reconstructed using the same filter, which design was based on modified 10-term cosine-sum window from [29]. The main requirements for this filtering operation is to keep $x(n)$ and

$y(n)$ in the audio band unchanged, while filtering out all artifacts above the audio band (designed filter has 10^{-13} dB passband ripple and about 300 dB stopband attenuation which result only from numerical errors). Designed low-pass filter's magnitude response and passband ripple is presented in Fig. 5.

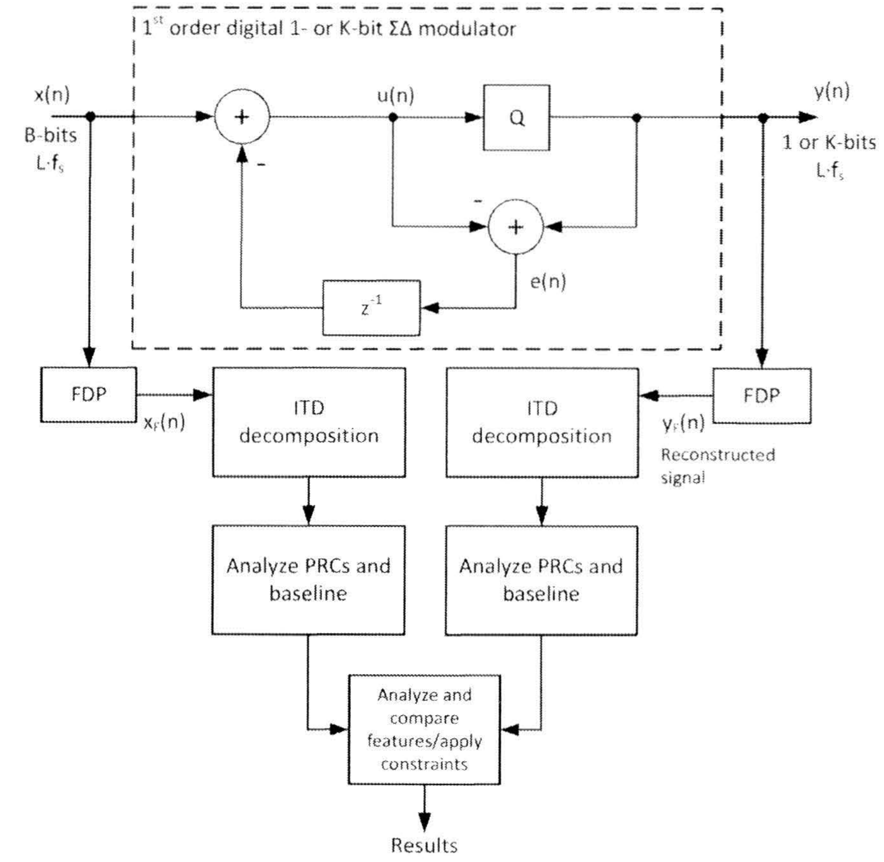

1st order digital 1- or K-bit ΣΔ modulator
x(n)
B-bits
L·fs
u(n)
Q
y(n)
1 or K-bits
L·fs
+
-
+
-
e(n)
z-1
FDP
xF(n)
ITD decomposition
ITD decomposition
yF(n)
Reconstructed signal
FDP
Analyze PRCs and baseline
Analyze PRCs and baseline
Analyze and compare features/apply constraints
Results

Figure 4. Block diagram of the proposed approach for analysis of one or multibit 1st order digital ΣΔ modulator with ITD method applied.

Since sample delay in digital ΣΔ modulator is placed in feedback path and low-pass filter has constant and known group delay, corresponding samples in signals $x_F(n)$ and $y_F(n)$ are perfectly aligned. ITD decomposes each signal into proper rotation components and baseline signals with temporal resolution equal to the time-scale of occurrence of extrema in $x_F(n)$ and $y_F(n)$. Thus, calculated instantaneous parameters or quantitative features from decomposed signals may be compared and analyzed without loss of short term (temporal) information embedded in input signals $x_F(n)$ and $y_F(n)$.

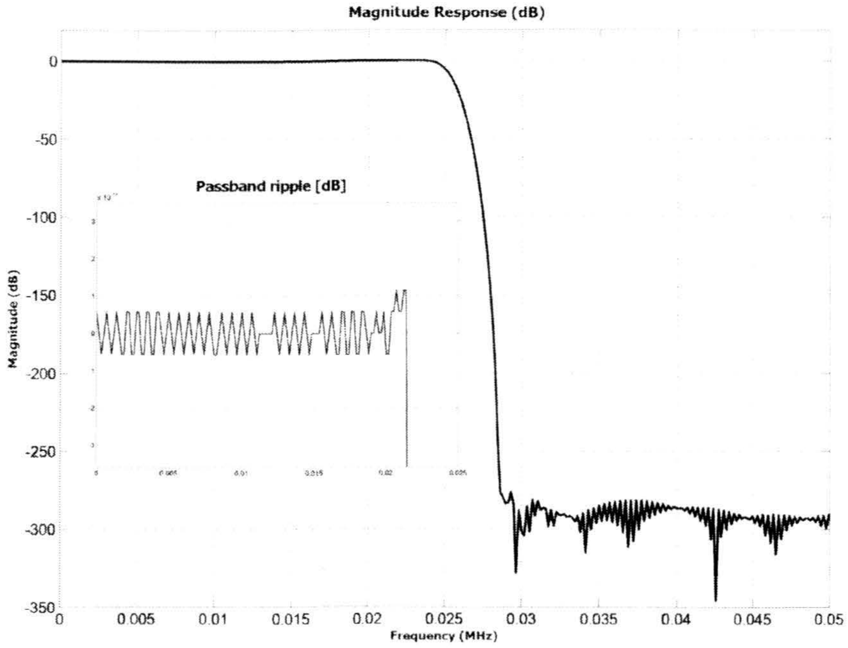


Figure 5. Magnitude response and bandwidth's ripple of designed low-pass filter with sampling frequency 64 × 44100 Hz (2.8224 MHz), about 300 dB stopband attenuation and 10^{-13} dB passband ripple.

V. Conclusions and Future Work

In this paper, the new approach for analyzing digital ΣΔ modulator's features is introduced. In particular, noise modulation phenomena is described more closely, because it is considered to result in deterioration in audio quality. According to [1-3,10,30] for high-quality audio applications the error should be ideally invariant with its input signal characteristics. Time-domain, frequency-domain and statistical-domain analyses proved that noise modulation exist in ΣΔ systems and showed how to deal with it. However, recent simulations were conducted with dc level or sinusoidal input signals to ΣΔ systems, which are not used in practical audio applications. Results from those simulations were estimated by averaging signals over time, thus short term (temporal) behavior of ΣΔ systems is missing. Author of this paper believes that results of the proposed analysis approach of digital ΣΔ modulators with real audio data as input signals will allow for more precise analysis of correlations between signals in such systems and to eliminate them. Verification of the proposed approach, reduction of possible problems that can disrupt results, simulations and careful analysis of results are the author's ongoing research.

References

[1] C. Dunn, M. Sandler, "Psychoacoustically optimal sigma-delta modulation", Journal of the Audio Engineering Society vol. 45 (4), pp. 212–223 (1997).

[2] S. P. Lipshitz, J. Vanderkooy, R. A. Wannamaker, "Minimally audible noise shaping", Journal of the Audio Engineering Society vol. 39 (11), pp. 836–852 (1991).

[3] J. Vanderkooy, S. P. Lipshitz, "Dither in digital audio", Journal of the Audio Engineering Society vol. 35 (12), pp. 966–975 (1987).

[4] C. Roads, "The computer music tutorial", 6th printing, Cambridge, MA: MIT Press (2002).

[5] B. Katz, "Dither", Digital Domain, FL (2007). <http://www.digido.com/bob-katz/dither.html>

[6] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of nonsubtractive dither", IEEE Trans. Signal Processing, vol. 48, no. 2, pp. 499-516 (Feb. 2000).

[7] R. A. Wannamaker(1992), "Psychoacoustically optimal noise shaping", J.Audio Eng. Soc., vol. 40, pp. 611-620 (1992).

[8] J. R. Stuart, "Coding high quality digital audio", Meridian Audio Ltd., UK, (Dec. 1997). <www.meridian-audio.com/ara/coding2.pdf>.

[9] C. R. Helmrich, M. Holters, U. Zölzer, "Improved psychoacoustic noise shaping for requantization of high-resolution digital audio", Audio Engineering Society 31st International Conference, London, UK (2007).

[10] M. Lewandowski, "Noise transfer function design and optimization for digital sigma-delta audio DAC", Archives of Acoustics, PAN, vol. 36, no. 1, pp. 87-108 (2011).

[11] C. Dunn, M. Sandler, "A comparison of dithered and chaotic sigma-delta modulators", Journal of the Audio Engineering Society vol. 44, pp. 227-244 (1996).

[12] S. P. Lipshitz, R. A. Wannamaker and J. Vanderkooy, "Quantization and dither: a theoretical survey", Journal of the Audio Engineering Society, vol. 40, pp. 355-375 (May 1992).

[13] J. D. Reiss, "Understanding sigma-delta modulation: the solved and unsolved issues", Journal of the Audio Engineering Society, vol. 56, no. 1 (2008).

[14] J. A. S. Angus, "Achieving effective dither in delta-sigma modulation systems", Audio Engineering Society 110th Convention, Amsterdam, Holland (May 2001).

[15] J. A. S. Angus, "Effective dither in high-order sigma-delta modulators", Audio Engineering Society 111th Convention, New York, USA (October 2001).

[16] J. Vanderkooy, S. P. Lipshitz, "Towards a better understanding of 1-bit sigma-delta modulators - part 3", Audio Engineering Society 112th Convention, Munich, Germany (May 2002).

[17] A. J. Magrath, M. B. Sandler, "Efficient dithering of sigma-delta modulators with adaptive bit flipping", Electronics Letters 31 (11) (1995).

[18] A. J. Magrath, M. B. Sandler, "Digital-domain dithering of sigma-delta modulators using bit flipping", Journal of the Audio Engineering Society vol. 45, no. 6 (1997).

[19] J. Reiss, M. B. Sandler, „The benefits of multibit chaotic sigma delta modulation", CHAOS 11 (2), 377 (2001).

[20] J. Reiss, M. B. Sandler, „Exploiting chaos in multibit sigma delta modulation", European Conference on Circuit Theory and Design, Espoo, Finland (August 2001).

[21] J. D. Reiss, M. B. Sandler „Multibit chaotic sigma delta modulation", Nonlinear Dynamics of Electronic Systems, Delft, The Netherlands, (June 2001).

[22] C. Dunn, M. B. Sandler, „A simulated comparison of dithered and chaotic sigma-delta modulators", Journal of the Audio Engineering Society vol. 44, no. 4 (1995).

[23] S. P. Lipshitz, J. Vanderkooy, „Towards a better understanding of 1-bit sigma-delta modulators - part 2", Audio Engineering Society 111th Convention, New York, USA (September 2001).

[24] J. D. Reiss, M. Sandler, „Dither and noise modulation in sigma delta modulators", Audio Engineering Society 115th Convention, New York, (October 2003).

[25] S. P. Lipshitz, J. Vanderkooy, "Dither myths and facts", Audio Engineering Society 117th Convention, San Francisco, USA (October 2004).

[26] E. Janssen, A. van Roermund, „Look-Ahead Based Sigma-Delta Modulation", Springer (2011).

[27] M. G. Frei, I. Osorio, "Intrinsic time-scale decomposition: time-frequency-energy analysis and real-time filtering of non-stationary sinals", Proceedings of The Royal Society A, 463, pp. 321-342 (2007).

[28] M. G. Frei, R. L. Davidchack, I. Osorio, "Least squares acceleration filtering for the estimation of derivatives and sharpness at extrema", IEEE Transactions on Biomedical Engineering 46, pp. 971-977 (1999).

[29] H. H. Albrecht, "A family of cosine-sum windows for high-resolution measurments", IEEE International Conference, ICASSP '01 Proceedings of the Acoustics, Speech, and Signal Processing, vol. 5, pp. 3081-3084, Washington, USA (2001).

[30] C. Dunn, "Psychoacoustic modeling of nonlinear errors", Measurment of Nonlinear Errors in Audio Electronics, Ph. D. Thesis, University of Essex, UK (1994).

Automatic Identification of Bird Species: a Comparison Between kNN and SOM Classifiers

Dorota Kamińska
Technical University of Lodz
Institute of Mechatronics and Information Systems
Poland, 90-924 Lodz
Email: kaminska.dorota@o2.pl

Artur Gmerek
Technical University of Lodz
Institute of Automatic Control
Poland, 90-924 Lodz
Email: artur.gmerek@p.lodz.pl

***Abstract* — This paper presents a system for automatic bird identification, which uses audio input. The experiments have been conducted on three groups of birds, which were created basing finishing on classification, the system is fully automated. The main problem in automatic bird recognition (ABR) is the choice of proper features and classifiers. Identification has been made using two classifiers - kNN (k Nearest Neighbor) and SOM (Self Organizing Maps). System has been tested using data extracted from natural environment.**

***Index Terms* — birds, kNN, HMM, recognition, identification, self organizing maps, SOM**

I. INTRODUCTION

The main goal of this paper was to develop an automatic system for bird recognition using audio input. This kind of system could be valuable for biological research and environmental monitoring. It is possible to recognize bird species using audio recording, basing on the fact that birdsongs have a grammatical structure and are composed of notes, syllables, phrases and calls (including alarm calls, distress calls, territorial calls and others). A set of one or more syllables and phrases arranged in a regular pattern is referred to as a song.

Classification of bird species by their sound is not a challenging task when they belong to different families. However, practical systems should be able to distinguish birds belonging to the same family but different species. Thus, experiments have been made on three different groups. The similarity between bird sounds in each group differ respectively: small, medium and significant difference. Groups have been created, based on correlation between the most descriptive features.

In order to chose proper methods of classification, literature study on whole spectrum of algorithms has been made. SOM and kNN classifiers, which gave satisfactory results, have been chosen and compared in this paper.

The majority of scientists who conduct research in this field use manual syllables division. Thus, currently existing systems are not fully automated. This paper presents a fully automated algorithm. Moreover, there is no problem with adjusting the system for new birds recognition using training module.

II. RELATED WORK

Analysis of bird sounds can be divided into three main parts: segmentation of bird sounds (e.g. to syllables), features extraction and classification. Scientists usually use manual or semi-automatic syllable segmentation. The last two stages of identification often differ greatly depending on individual approach.

Most of researchers use simple statistical features i.e. mean value, frequency bandwidth, duration of syllable, signal amplitude. Sometimes more sophisticated features are used e.g. Linear Predictive Coding (LPC), LPC Cepstral Coefficients (LPCCs), Mel-Frequency Cepstral Coefficients (MFCCs) [1] or wavelet coefficients.

Statistical classifiers like k nearest neighbors, bayesian classifiers and decision trees can be used for the purpose of bird recognition. Some methods, which are common for human voice identification like Dynamic Time Warping (DTW), Hidden Markov Models (HMMs) [2]–[4], Gaussian Mixture Models (GMM) and Vector Quantization (VQ) have been also used for birds species identification.

Lakshminarayanan et al. have introduced probabilistic models based on birdsong syllables [5]. Their Independent Frame Independent Syllable (IFIS) and Markov Chain Frame Independent Syllable (MCFIS) models achieved better results than Support Vector Machine (SVM) classifier.

Aki Harma has performed identification using sinusoidal modeling [6] basing on the fact that syllables can be approximated as varying amplitude and frequency brief sinusoidal pulses.

In recent years neural networks like Multilayer Perceptron (MLP) [7], Time Delay Neural Networks (TDNN) [8], Autoregressive Time-Delay Neural Networks (AR-TDNN) [9] and Self Organizing Maps [10] have been used by many scientists.

The most valuable are those publications, in which different methods are compared [11]. For example Briggs and others have presented a different statistical manifold approach [12].

McIlraith and Card have compared between backpropagation learning in two-layer perceptrons and discriminant analysis [13]. They have used simple statistical features (duration, mean, standard deviations, power spectral densities) and more complicated e.g. LPC. They achieved performance range from 82% to 93%, but experiments have been made merely on six different species.

III. Methods

Representation of the signal in time or frequency domain is a complex projection. Therefore features are sought to determine signal properties. In this study following features have been used: duration, bandwidth, fundamental frequency, power spectral density of a syllable and formant and antiformant frequencies. Also other features have been used, which are described in following subsections.

A. LPC coefficients

Linear predictive coding is a method used in audio signal processing for representation of spectral envelope of a digital signal in compressed form. Linear prediction, based on the assumption that a signal sample $u(n)$, can be approximated by linear combination of P previous samples for $n > 0$. The predicted signal value is expressed by the formula:

$$\tilde{u}(n) = -\sum_{p=1}^{P} a_p u(n-p)$$

$u(n-p)$ - previous observed values,
a_p - predictor coefficients,

LPC coefficients are determined by autocorrelation criterion. In this method the expected value of the squared error, which is defined as following equation, is minimized:

$$\sigma = E[err^2(n)] = \frac{1}{N-p}\sum_{n=p}^{N-1}\left[u(n) + \sum_{p=1}^{P} a_p u(n-p)\right]^2$$

where N is the number of samples.

To determine the optimal coefficients a_k, $1 \leq k \geq p$, a partial derivative of σ with respect to the variable a_p should be calculated and equated to zero. Afterwards p equations containing p variables are obtained with following solution:

$$Ra = -r$$

where R is a symmetric, autocorrelation matrix called Toeplitz matrix.

Experiments show, that the most optimal number of LPC coefficents is 12, therefore this amount has been used in this study.

B. Mel Frequency Cepstral Coefficients

Currently MFCCs are a standard in speech recognition [14]. The MFCC algorithm is multistage. At first, the signal is multiplied by the Hamming window, presented by the following equation:

$$w(n) \quad \begin{cases} 0.54 - 0.46\cos\left(\frac{2\Pi n}{N-1}\right) & , 0 \leq n \geq N-1 \\ 0 & , \text{otherwise} \end{cases}$$

where N specifies window size.
Subsequently FFT is computed. Then the estimation of power spectral density function is calculated and averaged using overlaping triangular weight functions. Design of the triangular functions includes *mel* scale. Following equations present the same frequency, m is frequency value in *mel*, f in *hertz* using natural logarithm:

$$m = 1127,01048\ln(1 + f/700), \quad f - 700(e^{\frac{m}{1127,01048}} - 1),$$

and using decadic logarithm:

$$m = 2595log(1 + f/700), \quad f = 700(10^{\frac{m}{2595}} - 1).$$

The last step is calculation of a Discrete Cosine Transform (*DCT*) of the logarithmed estimation, using the following formulas:

$$c_k = \sqrt{\frac{2}{L}}\sum_{l=0}^{L-1} ln\tilde{S}(l)cos\left(\frac{\Pi k}{L}(l + 1/2)\right), k = 0, 1, ...q - 1$$

where L is the number of weight functions and q is the number of Mel Coefficients.

IV. Classifiers

Classification is an algorithm, which assigns objects to groups (called classes) based on object features. Features are usually presented in a vector:

$$x_j = [x^1, x^2, x^3, ...x^d]$$

where d is the number of features, x^k is feature value.

All features values in a task are called the training set CU. It can be said, that the goal of classification is to assign a class $i \in M$ for an individual object x_j.

A. Nearest Neighbor Classifier (kNN)

In kNN algorithm the recognition process involves calculating distances in parameters space X between the unknown x_j object and all objects from the training set $x_k \in CU$ for $k = 1, 2, ..., I$, where I is the number of training examples.

In presented project euclidean distance has been used:

$$d(x_j, x_k) = \sqrt{\sum_{i=1}^{n}(x_{ij} - x_{ki})^2}$$

Obtained distances are sorted in an ascending order. Object x_j is assigned to this class, which is the most common among k nearest objects.

B. Self Organizing Map

Self Organizing Map is type of Artificial Neural Network (ANN). This type of ANN learns without a teacher, using only the observation of the input data (unsupervised learning). Network map, which creates a static grid cell, has a fixed size. It usually has a rectangular or hexagonal structure. Weights of input neurons can be initiated with random values. SOM has two basic methods of changing the neurons weights. The first one - Winner Takes All (WTA): the neuron, whose weights are closest to the input vector components is modified in such a way that its weights are as close as possible to theinput vector. The second one, Winner Takes Most (WTM):neuron with weight most similar to the input value is called the winner.

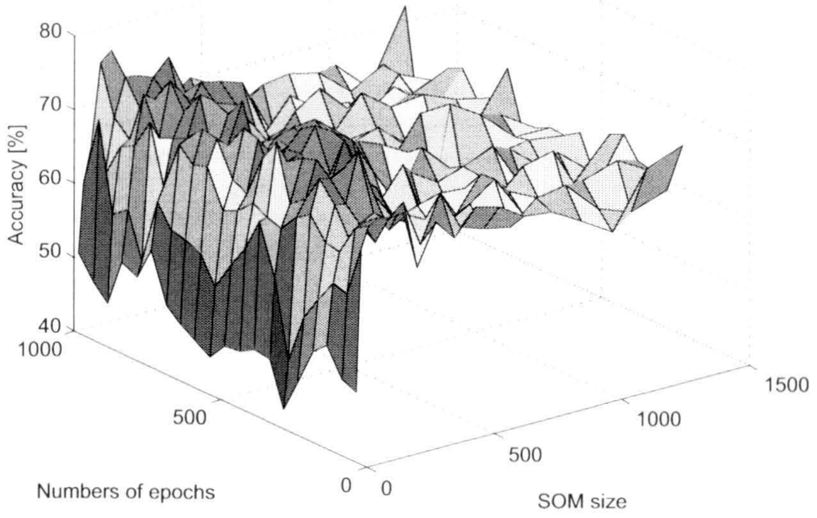


Fig. 1. The classification accuracy of birds sound with respect to the SOM size and number of epochs. The optimal number of neurons for given task was 450.

Its weights and neighboring neurons weight are modified. Frequently, this modification is dependent on the distance from the winner.

First step of the learning algorithm is to find the nearest maps element to the $c(x)$ vector.

$$c(x) = argmin||x - m_i||$$

where x is a sample from the training set CU in the step t

Then the winner and its neighbors are modified according to the formula:

$$m_i(t+1) = m_i(t) + h_{c(x),i}(t)[x(t)m_i(t)]$$

where $h_{c(x),i}$ is a neighborhood function given by:

$$h_{c(x),i} = \alpha(t)exp(-\frac{||r_i - r_{c(x)}||^2}{2\sigma^2(t)})$$

where $\alpha(t)$ is related to velocity of learning process.

Optimal number of learning epochs and SOM size has been calculated for this problem (Fig. 1). The network does not produce definite results of classification. It rather illustrates links between patterns by projecting them onto n-dimensional plane. After projection, data has to be decoded again in order to achieve accuracy of classification. This process is done by checking the distances between the nearest clusters of data around the point being the result of classification (Fig. 2).

V. EXPERIMENTS

Studies have been conducted on 10 birds species. All files have been downloaded from different Internet sources. The format of these files was PCM WAVE with 44100Hz sampling rate. 70% of files were used as training, 30% as testing set. Both sets were disjoint.

There are 3 groups of birds presented in the Table I. Species from these groups are correlated on a different level. For example the first group consists of birds, whose sounds are in high correlation (only from Passeriformes order). Thus, classification of birds from the first group could be more problematic in relation to other groups.

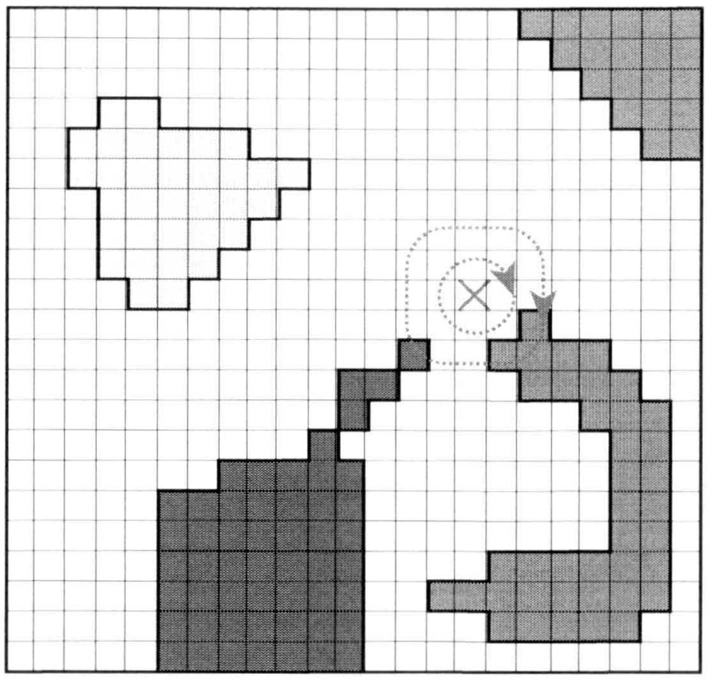


Fig. 2. An example of SOM processing. Colored areas represent different clusters (different bird species). Neighborhood urrounding results of classification (marked here as x) is measured based on euclidean distance. In a presented example result does not overlap with any areas, there are no clusters in the nearest distance of one checkered pattern. Because of that, algorithm checked the next checked patterns. There are 3 of them, which belongs to blue cluster and one, which represent to brown cluster. Consequently results will be classified as blue.

TABLE I
COMMON AND LATIN NAMES OF BIRD SPECIES OF THE BIRDSONG DATABASE AND THEIR CORRESPONDING AMOUNT OF SYLLABLES

Common name	Latin name	Training Syll.	Test Syll.
Great Tit	Parus major	562	137
Blackbird	Turdus merule	522	143
Eurasian Nuthatch	Sitta europea	530	104
Robin	Erithacus rubecula	543	140
Thrush Nightingale	Luscinia luscinia	370	80
Great Tit	Parus major	562	137
Blackbird	Turdus merule	522	143
Eurasian Nuthatch	Sitta europea	530	104
Grey Partridge	Perdix perdix	246	62
Tengmalm's Owl	Aegolius funereus	277	86
Common Swift	Apus apus	555	83
Wild Duck	Anas platyrhynchos	403	65
Common Cuckoo	Cuculus canorus	330	82
Grey Partridge	Perdix perdix	246	62
Tengmalm's Owl	Aegolius funereus	277	86

In order to prepare data for classification, signals were processed according to the algorithm presented in (Fig. 3). The first step - preprocessing, prepared the signal for features extraction. After that different features were calculated. The process of classification was divided into two stages: learning (teaching SOM classifier and creating a code book for kNN) and testing itself.

A. Preprocessing

The goal of preprocessing is adaptation and simplification of the signal for further analysis. It is divided into three steps:

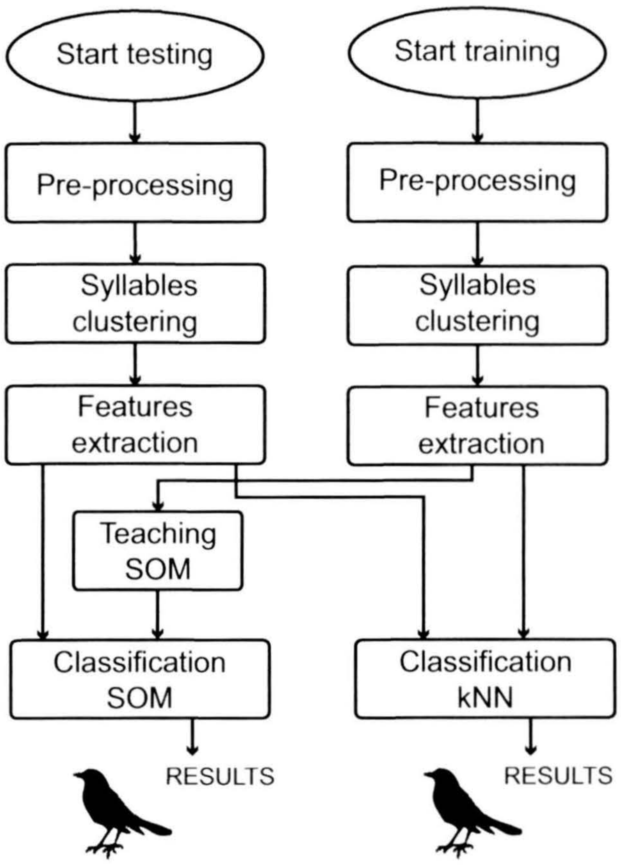


Fig. 3. Algorithm for processing audio files. Experiments have been conducted on 3 different groups of birds species.

filtration, normalization and wavelet decomposition. The aim of filtration, done by the use of band-pass filter, was to remove higher frequencies.

After filtration data were normalized. The goal of normalization was to eliminate the influence of the amplitude from the further analysis. Different amplitudes may be the result of various conditions during signal registration. In this study signal was normalized to fit $[-1, 1]$ value interval. Unfortunately normalization also decreased distances between classes. However, this was a necessary step, before proceeding to the next stages.

After normalization wavelet analysis was used for signal de-noising. Noise usually comes from recording apparatus, as well as from the environment. The first step of the method is decomposition. After selecting its level L and the type of wavelet functions, signal is divided into L decomposition levels according to the equation:

$$s(t) = \sum_{j=1}^{L} \sum_{k} d_j(k)\psi_j(t) + \sum_{k} c_L(k)\varphi_L(t)$$

where:
$\varphi_L(t)$ is a scaling function of the L-level,
$\psi_j(t)$ for $j = 1, 2...L$ are wavelet functions for the L levels.

B. Division into syllables

The definition of syllable is a problem in phonetics and phonology of human speech. This problem becomes even greater when it comes to birds. Therefore one of the biggest challenge of this study was syllables extraction. Physically, the syllable is defined as a segment which has higher intensity than its neighborhood. In this paper, following considerations are based on the signal time domain.

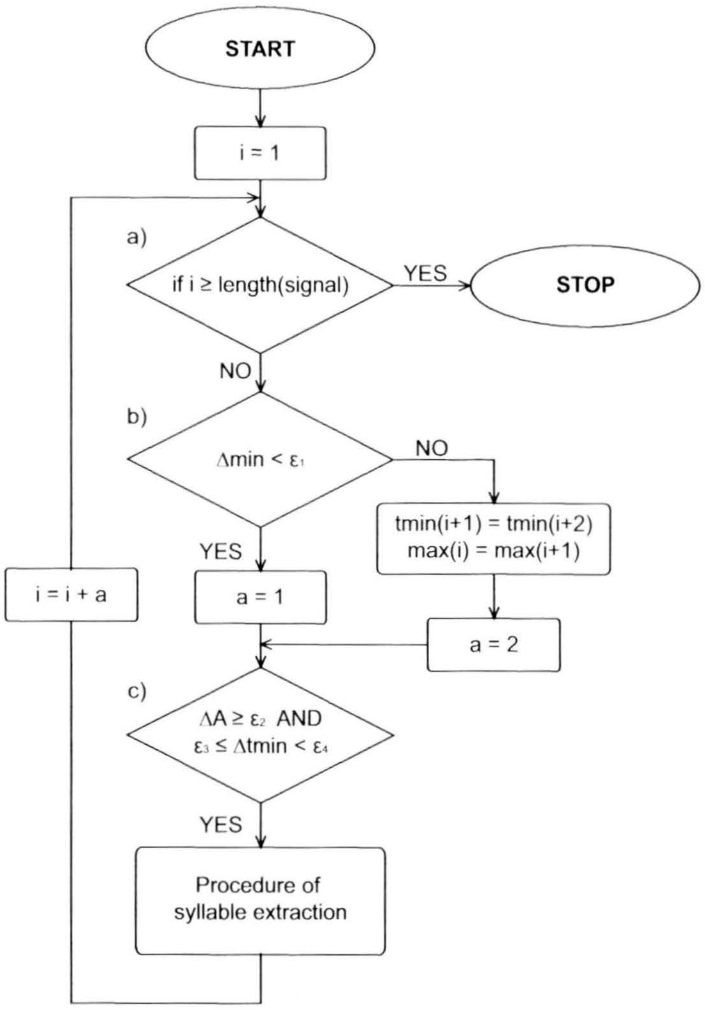


Fig. 4. Division into syllables algorithm

Division into syllables was divided into three parts. The first part was approximation, which reduced the noise and dimensionality of signal samples. After that local maxima and minima were designated, based on the gradient of signals polynomial approximation.

The syllables were clustered between two neighboring minima and usually had one maximum. However if a time period of a syllable was to small or differences between extrema were to low (what means, that this observation is a part of the same syllable), it was added to the previous syllable (Fig. 4).

Values of factors in this algorithm have great influence on classifiers performance. At the beginning the values of factors were established basing on the observation of the system, and after that, factors were optimized basing on the highest results of accuracy.

All the research and analysis was carried out on isolated syllables. Timing and syllable spectrogram are presented in the Fig. V-B.

After automatic division the features were extracted and clusters have been classified. During classification the methods described in the previous section have been used.

C. Results

The classification accuracy for different features shows that spectral features are the best for ABR task (Table II).

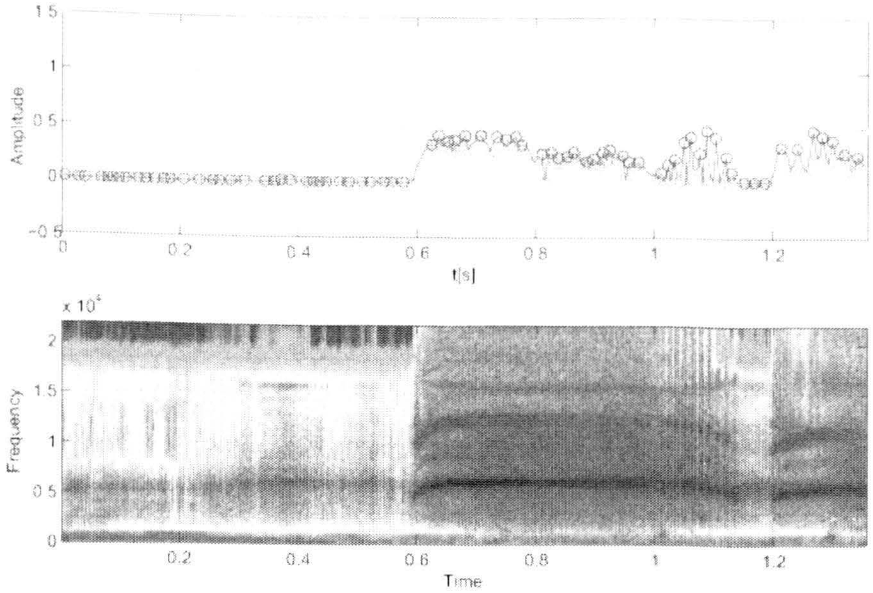


Fig. 5. (Upper Figure)Graph presents approximated signal (red line), as well as selection of minimal and maximal values (in circles). (Bottom Figure) Corresponding spectrogram of process signal.

TABLE II
CLASSIFICATION ACCURACY (CA) FOR SELECTED FEATURES

Feature name	CA(%)
PLP (Perceptual Linear Prediction)	79.89
LPC (Linear Predictive Coding)	74.6
MFCC (Mel Frequency Cepstral Coefficients)	80.42
Histogram	53.43
Formant	46.82
Antyformant	36.77
Bandwidth	26.19
Duration	23.28
FF (Fundamental Frequency)	27.78
PSD (Power Spectrum Density)	23.81
Sum of features	89.95

Table III presents the results of accuracy of classifiers for different birds species. One can observe that some birds species provide high classification accuracy, no matter to which group they were assigned to (Euroasian Nuthach, Tengmalms Owl). This means that the sound of those birds differ significantly from others in a particular group. One can also observe correlation between results accuracy and the ratio of training syllables to test syllables. At this point it is worth to note that a 30% of files were designated to test collection and not 30% of syllables. That is why the amount of bird syllables is different. Of course, if the number of training syllables was greater then test syllables, the accuracy could be higher. This regularity can be seen while comparing Eurasian Nuthatch (530 training syll., 104 test syll.) and Blackbird (522 training syll., 143 test syll).

VI. DISCUSSION

Achieved results are relatively satisfactory. It is difficult to exactly compare different works, because accuracy of classification depends greatly on the type of audio files and compared bird species. It is usually not a problem to identify birds, which sounds differ greatly, the problem is with similar bird sounds (e.g. from the 1st group). Results show that highest accuracy was achieved by the 3rd group, in which sounds of birds species differ emphatically.

TABLE III
CLASSIFICATION ACCURACY OF BIRDS SPECIES FOR 2 DIFFERENT CLASSIFIERS

Common name	CA_{kNN} (%)	CA_{SOM} (%)
Great Tit	59.85	43.79
Blackbird	43.36	30.07
Eurasian Nuthatch	70.19	61.54
Robin	31.43	15.00
Thrush Nightingale	57.5	21.25
Great Tit	59.12	37.96
Blackbird	44.06	36.36
Eurasian Nuthatch	73.08	59.62
Grey Partridge	66.12	40.32
Tengmalm's Owl	96.51	89.53
Common Swift	90.36	85.54
Wild Duck	93.85	73.85
Common Cuckoo	85.37	54.88
Grey Partridge	80.65	58.06
Tengmalm's Owl	97.67	86.05

There are unfortunately a few disadvantages of the system. First one is connected with the automatic syllables segmentation algorithm - the system has low immunity for various interferences. There may be a problem, when identifying bird sings on the same time with others. This problem can be considerable because birdsong is almost always connected with others (birds answer to each other).

Another problem is connected with values of various parameters in automatic syllables division algorithm. They were assign experimentally. One of the solution could be improving the algorithm by automatic adjustment of values of these coefficients basing on the information about expected group of recorded birds.

VII. CONCLUSIONS AND FUTURE WORK

In this article, the results of birds classification, based on their sounds have been presented. An automatic algorithm for division of bird sounds into syllables has been developed. Classification has been made using strictly selected features and 2 different classifiers. Tests have been made on real environment data sets. Mean accuracy of classification was 69,94 % for kNN and 52,92 % for SOM classifier. The highest accuracy has been achieved using MFCC features. The accuracy of classification depends mainly on the type of data sets, but also on used descriptors and classifiers.

Best results were achieved with kNN classifier. The research also shows that results are correlated with the similarity of the birds sounds. The experiments confirm that high accuracy in fully automated systems for ABR is possible, but not easy to achieve.

Future work will focus on adapting this system for handheld devices like cell phones or palmtops. These actions will be connected with optimalization of the algorithm in terms of speed of calculation and memory usage. Also a procedure, which will contrive with some difficulties mentioned in discussion section, should be developed.

New features, which are not connected to spectral construction of syllables, should be also tested. Descriptors from nonlinear dynamics, like fractal dimension or shapes of attractors can serve as an example of such features. Also additional features, extracted from phrases and songs, which show connections between syllables could be used.

ACKNOWLEDGMENT

This work is partly financially supported by the grant awarded to Artur Gmerek by the Ministry of Science and Higher Education of Poland (Grant No. N N514 469339).

REFERENCES

[1] C.-H. Chou, P.-H. Liu, and B. Cai, "On the studies of syllable segmentation and improving mfccs for automatic birdsong recognition," in *Proc. IEEE Asia-Pacific Services Computing Conf. APSCC '08*, 2008, pp. 745–750.

[2] T. S. Brandes, "Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1173–1180, 2008.

[3] C.-H. Chou, C.-H. Lee, and H.-W. Ni, "Bird species recognition by comparing the hmms of the syllables," in *Proc. Second Int. Conf. Innovative Computing, Information and Control ICICIC '07*, 2007, p. 143.

[4] E. Vilches, I. A. Escobar, E. E. Vallejo, and C. E. Taylor, "Targeting input data for acoustic bird species recognition using data mining and hmms," in *Proc. Seventh IEEE Int. Conf. Data Mining Workshops ICDM Workshops 2007*, 2007, pp. 513–518.

[5] B. Lakshminarayanan, R. Raich, and X. Fern, "A syllable-level probabilistic framework for bird species identification," in *Proc. Int. Conf. Machine Learning and Applications ICMLA '09*, 2009, pp. 53–59.

[6] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 5, 2003.

[7] C. E.D., "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," *Applied Acoustics*, vol. 62, pp. 1359-1374, 2001.

[8] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *Proc. 3rd Int. Conf. Intelligent Sensors, Sensor Networks and Information ISSNIP 2007*, 2007, pp. 293–298.

[9] S.-A. Selouani, M. Kardouchi, E. Hervet, and D. Roy, "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *Proc. ICSC Congress Computational Intelligence Methods and Applications*, 2005.

[10] P. Somervuo and A. Harma, "Analyzing bird song syllables on the self-organizing map," *Proceedings of the Workshop on Self-Organizing Maps (WSOM '03), Kitakyushu, Japan*, 2003.

[11] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895 – 2907, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/B6V15-49207P2-1/2/ef7fc86445321fbc2e001c70ff7e9f65

[12] F. Briggs, R. Raich, and X. Z. Fern, "Audio classification of bird species: A statistical manifold approach," in *Proc. Ninth IEEE Int. Conf. Data Mining ICDM '09*, 2009, pp. 51–60.

[13] A. L. McIlraith and H. C. Card, "A comparison of backpropagation and statistical classifiers for bird identification," in *Proc. Int Neural Networks,1997. Conf*, vol. 1, 1997, pp. 100–104.

[14] D. Niewiadomy and A. Pelikant, "Implementation of mfcc vector generation in classification context," *JOURNAL OF APPLIED COMPUTER SCIENCE*, 2008.

Estimation of Interaural Time Difference from Measured Head Related Impulse Responses

Michał Pec
Institute of Electronics,Technical University of Lodz,
Wolczanska 211/215, 90-924 Lodz, Poland
Email: michal.pec@gmail.com

Paweł Strumiłło
Institute of Electronics,Technical University of Lodz,
Wolczanska 211/215, 90-924 Lodz, Poland
Email: pawel.strumillo@p.lodz.pl

Abstract — **A novel approach to estimation of the Interaural Time Difference (ITD) from the measured Head Related Impulse Responses (HRIR) is proposed in the paper. An innovative application of the cross-correlation function that is estimated for impulse responses corresponding to adjacent sound arrival directions makes the presented method robust and immune to inherent noise components occuring in the measured HRIRs.**

I. Introduction

Head Related Impulse Responses are claimed to code all sound filtering phenomena responsible for spatial hearing in humans [1]. Because, however, HRIRs strongly depend on person's anatomy (pinnea, head and torso) individual characteristics need to be measured, typically in an anechoic chamber. The person under measurement has small microphones placed in each ear canals. Then wide-band sound sources are generated from different spatial locations surrounding the listener. Microphone responses to these sequentially generated sounds are recorded separately for left and right ear as $\mathrm{HRIR}_L(\theta, \varphi)$, $\mathrm{HRIR}_R(\theta, \varphi)$, i.e. impulse response functions parametrized in two angles defined in vertical-polar coordinate system Fig. 1, where θ is the azimuth and φ is the elevation angle. The widely used term Head Related Transfer Functions (HRTF) are the Fourier transforms of the corresponding HRIRs, and provide information on anatomy related spectral modifications of sound harmonics before they reach the listener's eardrums.

II. ITD estimation

HRIRs associated with a unique sound arrival direction are typically implemented as a cascade of a delay line and a minimum phase filter. In such a filtering scheme accurate estimation of the Interaural Time Difference is essential. ITD is the difference in arrival time of a sound wave between the left and the right ear. This time difference depends on the sound arrival direction identified by azimuth (θ) and elevation (φ) angles.

A number of methods for calculating of ITDs from measured HRIRs were proposed. In the interaural Cross-Correlation method the ITD is estimated as the lag of the maximum in the cross correlation function for impulse responses of the left and the right ear [4]. A computationally simpler approach is used in other methods, e.g. in [5], in which the sound arrival time delay is first estimated for each ear by comparing the corresponding HRIR samples to a predefined threshold (15% of maximum HRIR amplitude is typically used) and then the time difference between the two delays is taken as the ITD. ITD can also be computed from HRIR phase characteristic. Fitting a linear function to a phase of the HRIR or like reported in [6] to the excess phase (i.e. phase of all-pass component obtained from minimum phase reconstruction) is employed. However, the outlined approaches that are commonly used for ITD estimation may give wrong results under particular conditions. The correlation based methods can yield inaccurate estimation due to very different HRIRs shapes for the left and the right ear for some angles (e.g. due to sound shadowing by the listeners head at angles close to $\pm 90°$). Exemplary pair of HRIRs for source position $(\theta, \varphi) = (90°, 0°)$ is shown in Fig. 2. Maximum of cross-correlation function derived for such impulse responses may not indicate real time delay.

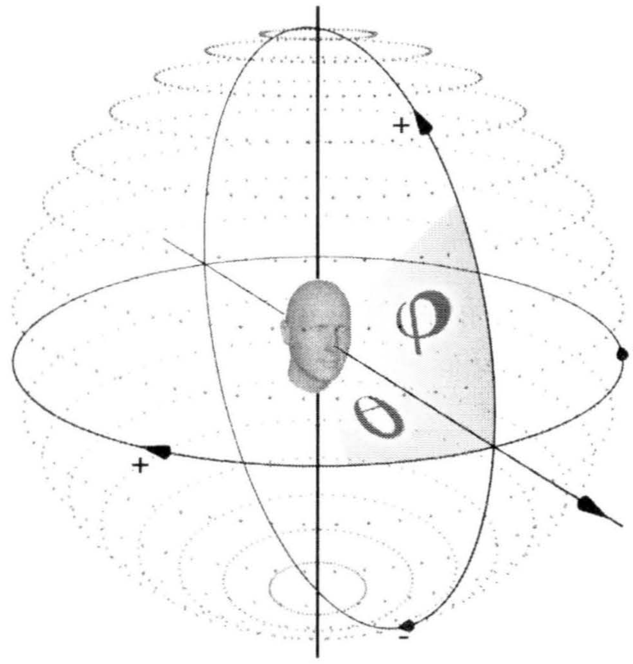


Fig. 1. A vertical-polar coordinate system

Although the acoustic conditions during the measurement are nearly anechoic, some reflections from elements of measurement apparatus or listener's body may influence recorded impulse responses. Such situation is shown in Fig. 3.

Applying threshold-based ITD computation method to such noised impulse responses may give unreliable estimations because the threshold value is crossed prematurely. On the other hand if such noise is removed using band-pass filtering, linearity of HRIR phase in particular frequency range may be

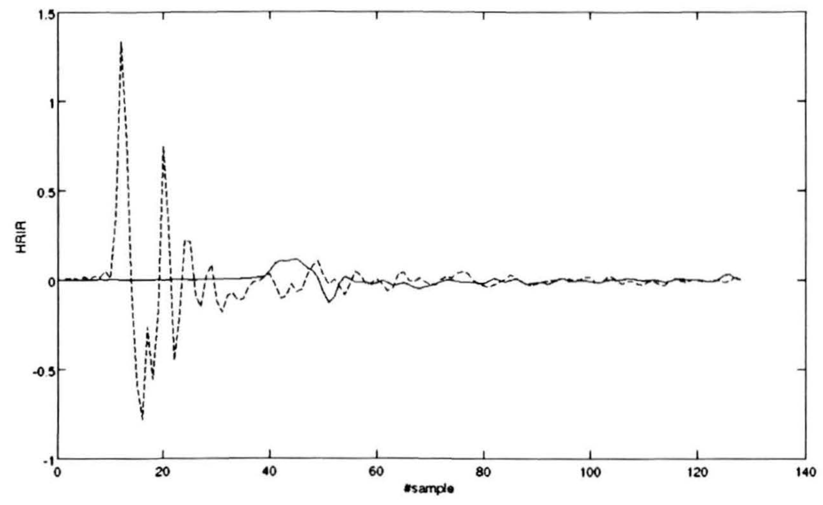


Fig. 2. HRIRs for left (solid line) and right (dashed line) ear for sound source position $(\theta, \varphi) = (90^\circ, 0^\circ)$; note a delay in the onsets of the impulse responses of approx. 0.6 ms ($fs = 44100\frac{1}{s}$)

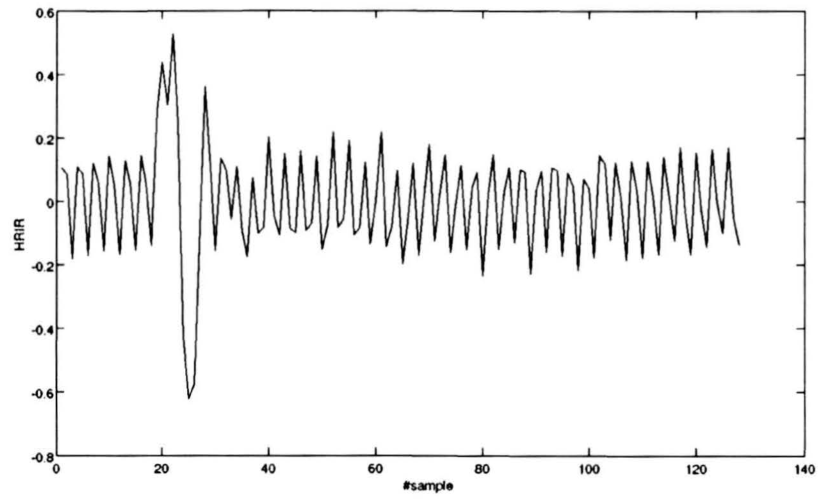


Fig. 3. Exemplary noised Head Related Impulse Response

lost which causes phase-based ITD estimation methods fail.

Values of ITD can be also calculated basing on the anthropometrical features using the so called Woodworth's formula [7]. In this approach to ITD estimation, a simplified (spherical or ellipsoidal) model of listener's head is used to calculate the difference of the paths lengths from a sound source to left and right ear correspondingly. This method, however, requires conduction of precise measurements of the anthropometric features of the listener. Consequently, the HRIR acoustic measurements must by accompanied by additional non-acoustic measurements which lengthen the procedure and can be a source of additional errors in ITD estimation. Authors experienced the indicated problems with ITD estimation from HRIRs, thus decided to search for a more robust and accurate ITD estimation method.

III. The proposed ITD estimation method

The proposed ITD estimation method employs the cross-correlation function of the HRIRs which is computed for the same ear and the adjacent sound arrival directions. In such a way a map of elementary time differences can be built. The Elementary Time Difference (ETD) is defined here as the difference in sound wave's arrival times for two adjacent elevations. By appropriately summing up consecutive ETDs, single ear time delays for any direction can be obtained. Individual ITDs can be calculated by subtracting the right-ear time delay from the corresponding left-ear time delay. The advantage of this simple, yet effective, ITD estimation method is that the cross-correlation function is estimated for impulse responses of similar shape waveforms. Exemplary pairs of HRIRs for adjacent sound source's positions for left and right ear correspondingly are shown in Figs. 4 and 5.

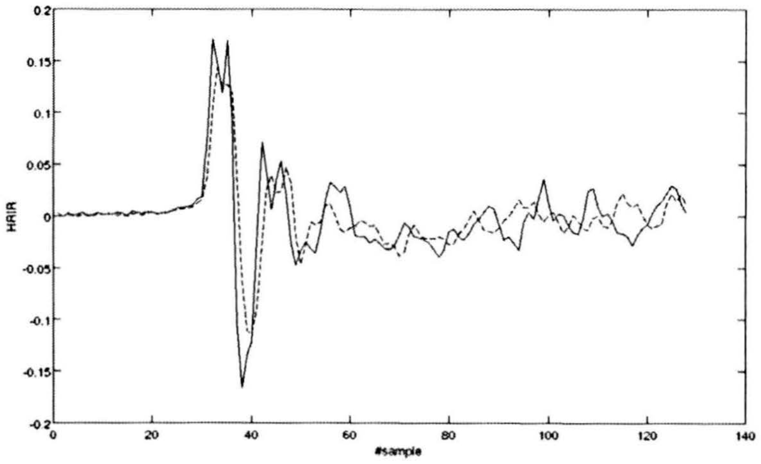


Fig. 4. HRIRs for left ear for two sound source's positions: $(\theta, \varphi) = (60^\circ, 45^\circ)$ - solid line - and $(\theta, \varphi) = (60^\circ, 36^\circ)$ - dashed line

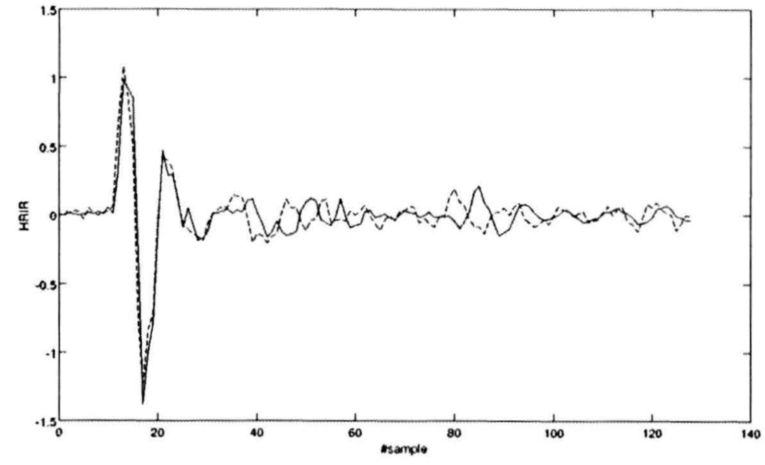


Fig. 5. HRIRs for right ear for two sound source's positions: $(\theta, \varphi) = (60^\circ, 45^\circ)$ - solid line - and $(\theta, \varphi) = (60^\circ, 36^\circ)$ - dashed line

A method based on a similar scheme was proposed in [8], but in our opinion the earlier reported method is more complicated computationally and insufficiently documented. The HRIR set for which the ITD is to be calculated should fulfill the two requirements:

- directions for which the HRIRs are measured are given in a grid defined in a vertical-polar coordinate system shown in Fig. 1. Two angles: azimuth(θ) and elevation(φ) specify a sound source's angular location.
- the HRIRs are measured for elevation angle $\varphi = 90^\circ$ (straight above the head) for which the ITD is assumed to be 0.

Block diagrams of the proposed algorithm for ITD estimation are shown in Figs. 6 and 7.

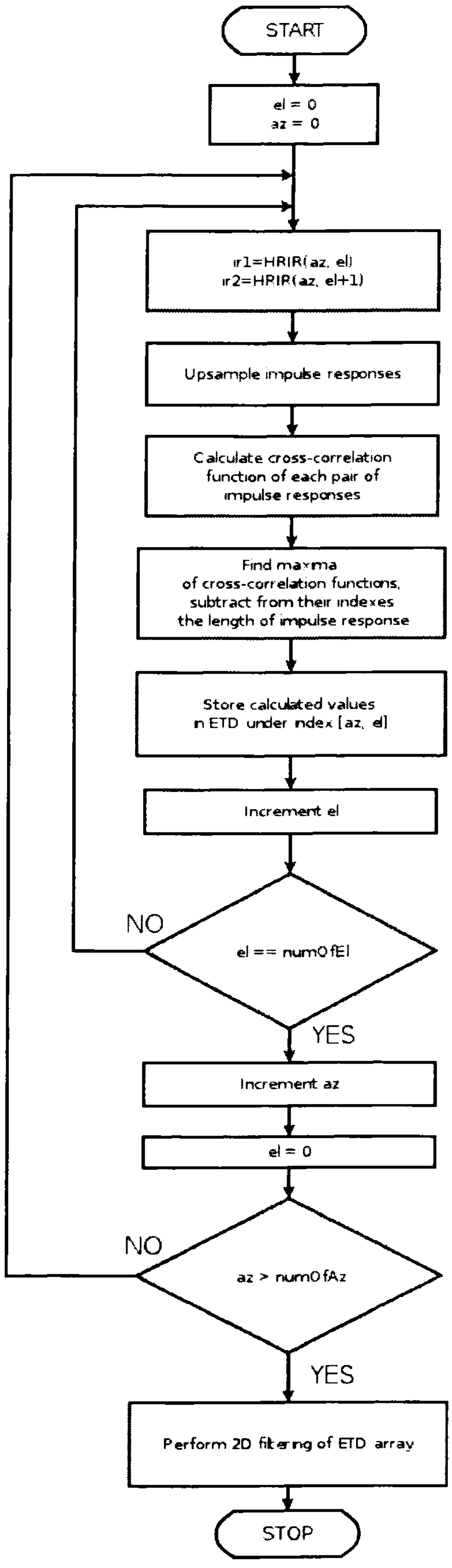

START
el = 0
az = 0
ir1=HRIR(az, el)
ir2=HRIR(az, el+1)
Upsample impulse responses
Calculate cross-correlation function of each pair of impulse responses
Find maxima of cross-correlation functions, subtract from their indexes the length of impulse response
Store calculated values in ETD under index [az, el]
Increment el
NO
el == numOfEl
YES
Increment az
el = 0
NO
az > numOfAz
YES
Perform 2D filtering of ETD array
STOP

Fig. 6. ITD calculation algorithm - part I

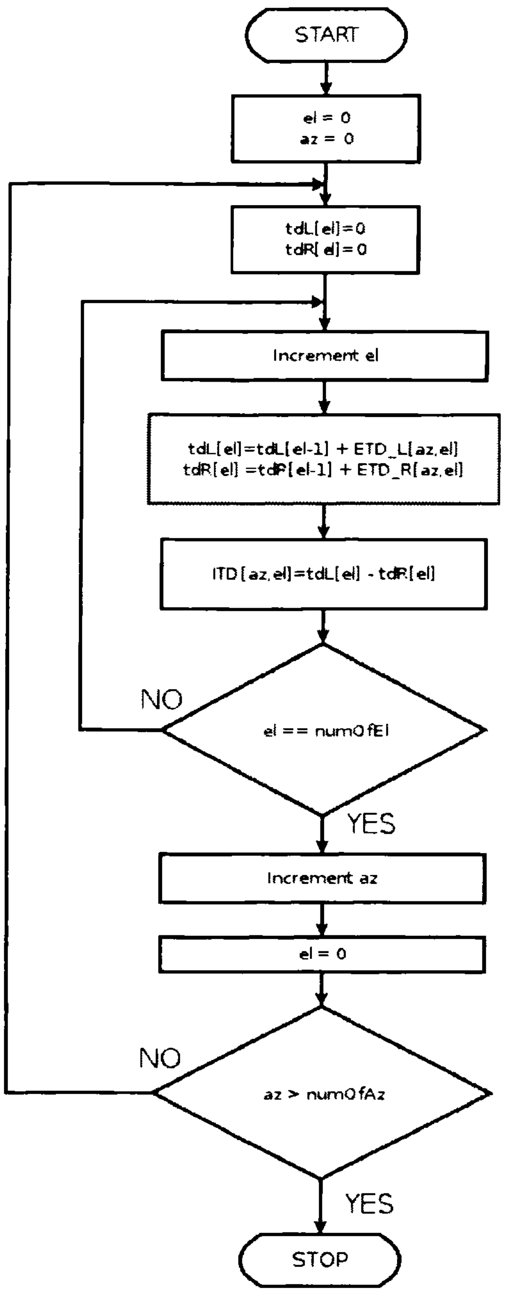


Fig. 7. ITD calculation algorithm - part II

TABLE I
2D LOW-PASS FILTER COEFFICIENTS

0.0357	0.0357	0.0357
0.0357	0.7143	0.0357
0.0357	0.0357	0.0357

For a clearer explanation its description is divided up into the two following parts:

- computation and processing of ETD
- computation of Interaural Time Difference

Variables: *el* and *az* in the diagrams are actually integer indices of consecutive elevation and azimuth angles. The only exception from this scheme occurs in Fig. 6 in the third topmost block where impulse responses are being picked to analysis. Since HRIR is a function of azimuth and elevation angles, in this operation *el* and *az* act as parameters of a function, which is denoted by round brackets. Elevation index $el = 0$ corresponds to source position stright above listener's head.

Up-sampling applied to impulse responses ensures an improved resolution of the estimated ETDs. In the reported method, correct estimation of each ETD is critical. Every single error made in calculating the cross correlation affects the final ITD value for the successive elevations. Low-pass filtering applied to 2D arrays of ETDs is not necessary, but significantly reduces the influences of possible estimation errors due to noisy or interfered impulse responses. On the other hand, filter transition band must be not too sharp, because it may influence the value of ITD. Filter coefficients used for EDT smoothing are shown in Table I.

IV. Results

Validity of the proposed ITD estimation approach was evaluated by comparing it to other earlier proposed methods. HRIR sets for which ITDs were estimated were measured in an anechoic chamber using the system owned by the Technical University of Lodz. The measurement set-up and the adopted procedure were described in a more detail in [3]. HRIR recordings were performed with $\Delta\theta = 10^\circ$ azimuth and $\Delta\varphi = 9^\circ$ elevation resolution. The two following methods were used for comparing earlier reported ITD estimation techniques to the one proposed in this work:

- Comparison of smoothness of 2D surfaces of ITDs indexed by azimuths and elevations in graphical and numerical way - this allows to evaluate method robustness against interferences occurring in the measured HRIRs.
- Comparison of ITD values derived with proposed and earlier reported methods.

A. Smoothness analysis

Figures 8 - 11 show values of ITDs for a sample HRIR set estimated with the compared methods. ITD is shown in a form of 2D mesh plot indexed with azimuth and elevations.

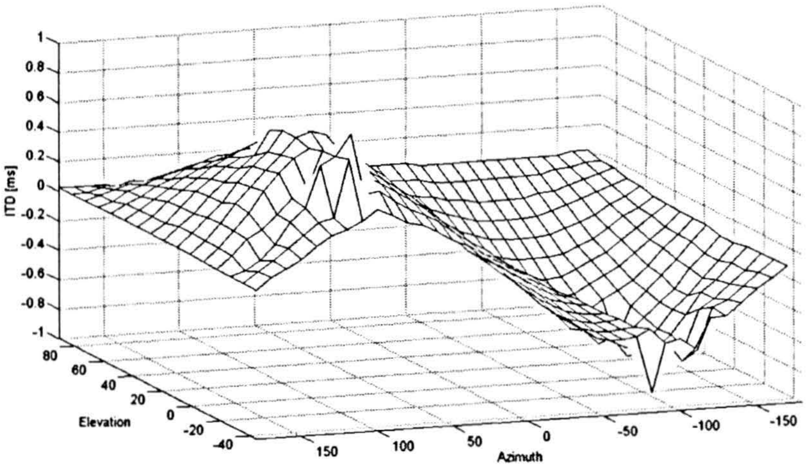


Fig. 8. ITD estimated using the interaural cross-correlation method (elevation and azimuth angles are given in degrees)

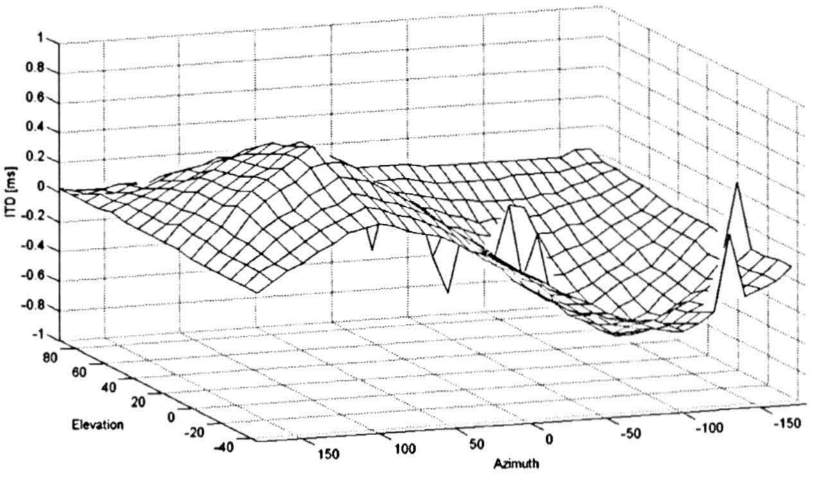


Fig. 9. ITD estimated using the impulse response thresholding

ITD estimation results obtained by the interaural cross-correlation method are shown in Fig. 8. Note a number of larger variations in the estimated ITD values for azimuths close to -90° and 90° where the HRIRs for left and right ear significantly differ due to the shadowing of one of the ears. In Fig. 9 , which shows results obtained for the impulse response thresholding method, a number of incidental errors occur. These errors might be due to noise components in the measured HRIRs. A few errors occur also in Fig. 10 , in which the ITD values are estimated by means of linear HRIR phase fit over a frequency range from 500 to 2000Hz. Note that ITD estimation results obtained by using the proposed method yield the most even surface in comparison to the earlier proposed methods (Fig.11).

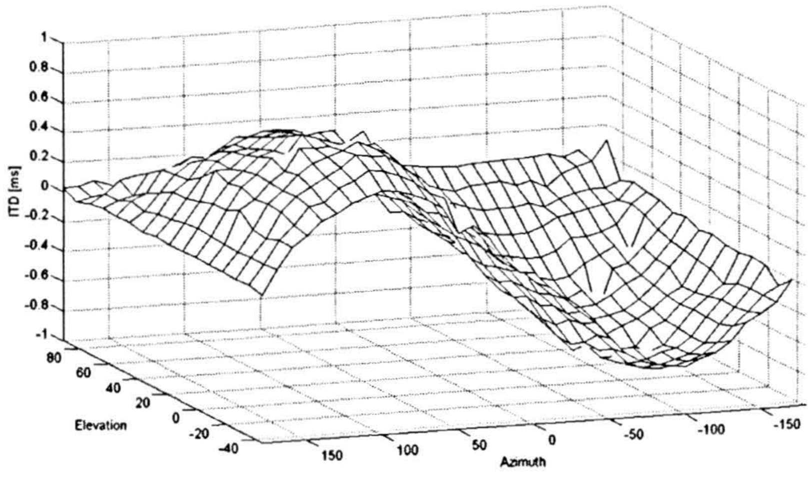


Fig. 10. ITD estimated by the linear HRIR phase fit method

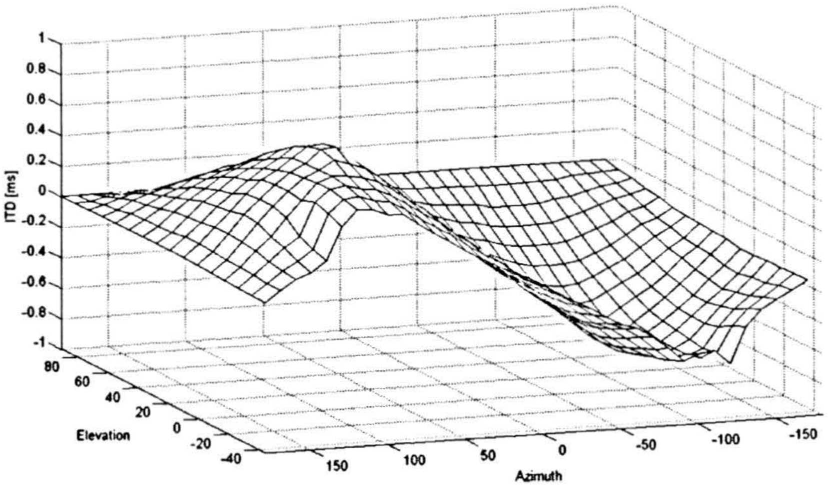


Fig. 11. ITD estimated using the proposed method

In order to allow for objective comparison of smoothness of ITD we propose the smoothness measure basing on energy of the first derivate in azimuth and elevation domain. The proposed measure is given by the following equation:

$$Sm = \sum_{el}\sum_{az} ITD_{az}^{'2} + \sum_{az}\sum_{el} ITD_{el}^{'2} \qquad (1)$$

where $ITD_{el}^{'}$ and $ITD_{az}^{'}$ are the gradients elevation- and azimuth-wise of the 2D array of ITD values. Table shows the values of such measure calculated by means of the discussed methods for three exemplary HRIR sets. The smoothness parameter was calculated for ITD given in miliseconds. The lower is the value of this measure the smoother the data is.

TABLE II
COMPARISON OF SMOOTHNESS OF ITDS ESTIMATED BY APPLYING DIFFERENT METHODS

HRIR set	Smoothness (the proposed method)	Smoothness (interaural cross-correlation method) [4]	Smoothness (threshold detection method) [5]	Smoothness (linear phase fit method) [6]
1	3.21	4.51	4.78	5.4
2	3.81	4.62	4.55	4.6
3	3.15	3.98	4.45	4.25

TABLE III
VALUES OF AVERAGED DIFFERENCE OF ITD VALUES, STANDARD DEVIATION OF DIFFERENCE AND ROOT MEAN SQUARED ERROR FOR THREE HRIR SETS

HRIR set	Average difference of ITD [ms]	Standard deviation of difference	Root Mean Square error
1	0.016	0.033	0.037
2	-0.006	0.023	0.022
3	0.01	0.026	0.033

B. Comparison of ITD values

Values of the Interaural Time Difference calculated by means of the proposed method were compared to values of ITD obtained using cross-correlation, thresholding and linear phase fit methods. Since ITDs given by those methods may contain incidental errors due to phenomena mentioned in section II, the authors have decided to use median of ITD calculated with classical methods for all directions as a reference to test new method. Performance of the proposed method was evaluated by means of the following quality measures:

- averaged difference of ITD values

$$\bar{E} = \frac{1}{N}\sum_{i=1}^{N}(ITD_{ref}(i) - ITD_{new}(i)) \qquad (2)$$

- standard deviations of differences

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Big((ITD_{ref}(i) - ITD_{new}(i)) - \bar{E}\Big)^2} \qquad (3)$$

- root mean square error

$$RMSe = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(ITD_{ref}(i) - ITD_{new}(i))^2} \qquad (4)$$

Values of these quality measures calculated for the considered three HRIR sets are shown in table III.

Values shown in table III indicate some discrepancies between ITD values derived with the proposed method and the previously reported ones. However, these differences should not be alarming, if we keep in mind the susceptibility of the earlier documented methods to noise and other interferences. Deviation of results obtained with the proposed method should not exceed 2 samples for a sampling rate $fs = 44100\frac{1}{s}$ that was used for the HRIR measurements.

V. CONCLUSIONS

A novel approach to estimation of Interaural Time Differences that is important in spatial sound reproducing techniques was proposed. The method reported here outperforms other commonly used methods in terms of computation simplicity and robustness against incidental interferences and noise components present in the measured HRIRs.

ACKNOWLEDGMENT

The project is in part financed by the National Center of Research and Development in years 2010-2013 under the grant NR02-0083-10.

REFERENCES

[1] B. C. J. Moore, *Introduction to psychology of hearing*, Academic Press Inc., 2004

[2] C. Cheng, G.H. Wakefieldk, Introduction to head-related transfer functions (HRTFs): representations of HRTFs in time, frequency, and space, *J. Acoust. Soc. Amer.*, 49, 2001, pp. 231–219

[3] A. Dobrucki, P. Plaskota, P. Pruchnicki, M. Pec, M. Bujacz and P. Strumillo, Measurement system for personalized head-related transfer functions and its verification by virtual source localization trials with visually impaired and sighted individuals, *J. Acoust.Soc. Amer.*, 58, 2010, pp. 724–738

[4] D.J. Kistler, F.L. Wightman, A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction, *J. Acoust.Soc. Amer.*, 91, 1992, pp. 1637–1647

[5] R. Duda, W. Martens, Range dependence of the response of a spherical head model,*J. Acoust.Soc. Amer.*, 104, 1998 pp. 3048–3058

[6] J. Huopaniemi, J.O. Smith, Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters, *16th International Conf. of the Audio Eng. Soc.*, Rovaniemi, Finland, 1999

[7] R.O. Duda, C. Avendano, V.R. Algazi, An adaptable ellipsoidal head model for the interaural time diffrence. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 99*, 1999, pp. 965–968

[8] M. Toyoda, K. Watanabe, Y. Iwaya, Y. Suzuki, An estimation method of interaural time differences from measured head-related impulse responses, *Acoustical Science and Technology*, 27, 2006, pp. 239–241

SESSION 3:

AUDIO PROCESSING II

Hybrid Sinusoidal Modeling of Music with Near Transparent Audio Quality

Maciej Bartkowiak, Łukasz Januszkiewicz
Chair of Multimedia Telecommunications and Microelectronics
Poznan University of Technology
Poznan, Poland
mbartkow@multimedia.edu.pl

Abstract — **It is often believed that sinusoidal as well as sinusoidal plus noise modeling is not capable of delivering high audio quality for complex signals such as wideband music. We identify the key sources of modeling artifacts in sinusoidal modeling systems and demonstrate a hybrid system that offers near transparent quality of reconstructed audio, thanks to application of a dedicated transient model, an accurate parameter estimation method, an advanced tracking algorithm and a warpedfrequency spectral model of noise.**

Keywords — ***audio analysis/synthesis; sinusoidal model; estimation; transient modeling; noise modeling***

I. Introduction

Sinusoidal model (SM) and hybrid sinusoidal + noise model (HSNM) are both well-established frameworks for signal analysis, transformation and synthesis, as well as enhancement, source separation, recognition, and data compression [1]. Since its early introduction in late 1980s, this family of techniques has been applied mostly for representing speech and single instrumental sounds [1,2,3]. Still, there is a common belief that such representation does not offer a high quality reconstructed audio for wideband signals, especially in challenging cases, like e.g. complex music with many sound sources of wide dynamics and spectra. In this paper we discuss and show a solution to several limitations of HSNM. We also present an advanced modeling system that offers a near transparent audio quality, i.e. the reconstructed signal is in most cases perceptually undistinguishable from the original audio, while a compact and meaningful parametric representation is achieved, enabling efficient implementations of auditory scene analysis, transformations, and data compression.

Generally speaking, SM consists of several stages, including spectral analysis, detection of spectral peaks identifying sinusoidal components, parameter estimation of sinusoids, and tracking of those parameters across consecutive audio frames. In this approach, all signal components are represented by modulated sinusoids, although it may be inefficient for signals with a significant amount of noise. Hybrid sinusoidal modeling addresses this problem by introducing additional modeling components. However, it requires a separation of the original signal into sinusoidal part and non-sinusoidal part, which may be particularly difficult for complex audio. In typical implementations, only certain peaks of short time spectrum are identified as sinusoids (the deterministic component), tracked, and synthesized. The residual from the sinusoidal part represents the remaining spectral energy. This residual (noise component) may be modeled as an auto-regressive random process characterized by a time-variable spectral density function and a temporal magnitude envelope [3,4],

$$\hat{x}(t) = \underbrace{\sum_{k=1}^{K} A_k(t) \sin\left(\varphi_k + 2\pi \int_0^t f_k(\tau)\, d\tau \right)}_{\text{deterministic component}} + \underbrace{h_n(t) * \xi(t)}_{\text{noise component}} . \qquad (1)$$

II. Artifact sources in HSNM

Sinusoidal analysis aimed at detection of spectral peaks representing all important tonal components is usually implemented on a frame basis, as a short-time Fourier transform (STFT) followed by picking salient peaks of magnitude spectrum. STFT-based sinusoidal analysis is always a compromised solution, trading off the accuracy of representing modulated partials for the ability to capture all low-frequency partials, which is more important since they usually describe fundamental harmonics of many musical sounds and exhibit high energy. The problems with STFT are its fixed spectro-temporal resolution as well as the underlying assumption on local stationarity. High spectral resolution required for proper analysis of low pitched sounds (sometimes below 50Hz) enforces the use of long analysis windows (100-200ms, i.e. 2^{12}-2^{13} samples if $f_s = 44.1$kHz) in order to reliably resolve individual harmonic partials. Higher frequency components in wideband audio often exhibit deep frequency and amplitude modulations, and they must not be considered stationary within a time window of such length. On the other hand, relying on STFT with long and strongly overlapping analysis windows usually yields significant pre-echo artifacts when analyzing sounds with transients.

Transients are relatively sparse, but very important elements of sounds characterized by sudden increase of energy. Many forms of spectral processing of audio have insufficient temporal resolution that results in temporal smearing of transients, which is easily detectable and usually annoying for the listener. Since there is no way to effectively estimate transients with highly overlapping STFT frames, it may be concluded, that a separate model of transients [6] with transient removal before sinusoidal analysis, as well as transient-aware multi-resolution sinusoidal analysis, are both necessary for high quality representation.

In a typical HSNM system, only spectral peaks representing actual sinusoids should be selected for the tonal part of the model, and their parameters should be tracked in order to establish sinusoidal trajectories. In practice, discerning between sinusoidal and stochastic spectral peaks is a challenging problem. First of all, the bulk of spectral components observed in natural audio is neither purely sinusoidal nor purely random. Several techniques for classification of spectral peaks have been proposed [7,8], but the general experience is that applying such selection is always prone to misclassification and audible modeling errors. In our experience, the best verification of whether given spectral peak is a sinusoidal one, is if it yields a reliable knot of a sinusoidal trajectory as a result of tracking. Therefore, application of any peak classification criteria should be very conservative in order to avoid rejecting weak sinusoidal partials which may be obscured by noise.

Accurate partial tracking is probably the most challenging problem in HSNM, because the goal is not well defined and it depends on particular application. For example, too fragmented trajectories resulting from too conservative connection rules yield a model that is inefficient in terms of data compression. Conversely, long and continuous trajectories obtained by excessive linking of partial data representing actually different sources may result in significant errors in source separation. In our experience, simple tracking algorithms [2,3] are inappropriate for modeling of wideband music because of not taking into account the wider temporal context of established connections and because of too simplified connection criteria, depending mostly on absolute frequency difference and not reflecting deeper modulations observed in upper harmonics. Trajectories obtained from a simple tracking algorithm are usually fragmented and chaotic. A signal synthesized from such a model is inferior in quality due to many audible discontinuities of partials which cannot be easily masked by applying a smooth fade-in and fade-out to segments at trajectory endpoints, since such amplitude modulation introduces a significant spectral distortion. For the sake of preserving the continuity as much as possible, tracking based on various forms of adaptive prediction is preferable in high quality HSNM.

Handling the non-sinusoidal component by a separate model requires obtaining the residual of SM as clean and free from unwanted sinusoids as possible, because otherwise the residual spectral model tends to compensate for their energy, and the amount of noise becomes overestimated. The residual may be derived as a time-domain difference between the original signal and the synthesized sinusoids [2], or through spectral subtraction [3]. The first option requires an accurate estimation of partial parameters, as well as phase-coherent synthesis. The latter option is more tolerant to estimation errors, but it usually yields the power density spectrum being underestimated.

Spectral modeling of the residual, interpreted as a random noise, is often implemented in a form of auto-regressive modeling, or linear prediction (LP). Unfortunately, this popular technique [1,3,4] is not well suited for modeling colored noise components in wideband music, because of its frequency resolution being uniform in a linear scale which does not match the non-uniform resolution of human ear. Hence, an LP model of reasonable order is very inaccurate in the low frequency range while it is unnecessarily accurate in high frequencies.

III. SinMod toolbox

A hybrid sinusoidal modeling system has been developed for dealing with wideband complex music signals in the HSNM framework. The software implementation in a form of a Matlab toolbox is freely available for non-commercial applications at http://www.multimedia.edu.pl/audio_research/.

It has been verified through a number of blind listening experiments, that this system offers a near transparent quality, i.e. the reconstructed audio is perceptually nearly undistinguishable from the original music recording. The key elements that contribute to this high fidelity are:

- a dedicated transient model, with transients re-synthesized and removed from the signal prior to sinusoidal analysis,
- multi-resolution sinusoidal analysis for detection of both low frequency dense partials and deeply modulated higher frequency partials,
- adaptive prediction-based partial tracking for creating long, continuous and meaningful trajectories,
- post-processing of sinusoidal trajectories to cope with overestimation and fragmentation of trajectories,
- accurate sinusoidal parameter re-estimation once tracks are established, enabling accurate and phase-coherent synthesis,
- a noise model with frequency resolution corresponding to the resolution of human ear.

These elements will be discussed in the remaining part of this paper.

A. *Modeling of transients*

Before transient modeling, a reliable detection is to be performed. Popular transient detection techniques are based on thresholding of certain audio features, like local energy or spectral flux. In the HSNM system proposed here, a complex spectral domain prediction (2) is employed [9] for detecting sudden changes of signal short-time amplitude and phase spectrum, usually associated with discontinuities, note onsets, or short bursts of energy accompanying transients,

$$\eta(m)=\sum_{k-1}^{K}\left|X_k(m)-\hat{X}_k(m)\right|, \qquad (2)$$

where $\hat{X}_k(m)=\left|X_k(m-1)\right|\exp\left[j2\varphi_k(m-1)-j\varphi_k(m-2)\right]$ is a complex-valued prediction of a DFT co-efficient $X_k(m)$ based on two previous frames, m-1, and m-2.

The detection function $\eta(m)$ as proposed in [9] is correlated with signal magnitude, therefore an adaptation to local signal dynamics is necessary for reliable detection. A decision process with an adaptive hysteresis is used here. The lower and upper thresholds are dependent on local mean and median values of $\eta(m)$. Furthermore, in order to avoid false alarms on pure noise, transient detection is enabled only when signal amplitude exhibits a significant local peak.

For transient modeling, a simple model of damped sinusoids sharing a common amplitude envelope is adopted from [10]. In the first step, a parameterized envelope model (so called Meixner function) is fitted to the magnitude of the signal within a short rectangular window. Subsequently, a set of sinusoidal modulating components is iteratively detected based on FFT analysis of the original signal windowed by the envelope determined in the first step. Finally, the phase of each sinusoid is estimated using a least-squares fit.

The procedure results in a set of data consisting of three envelope parameters, identifying the position, attack time and decay time of each transient, as well as frequencies, amplitudes and phase of each of the modulating sinusoids. The signal synthesized from these parameters matches the original waveform and may be subtracted in time domain resulting in a conditioned signal that is better suited for sinusoidal analysis (cf fig. 1).

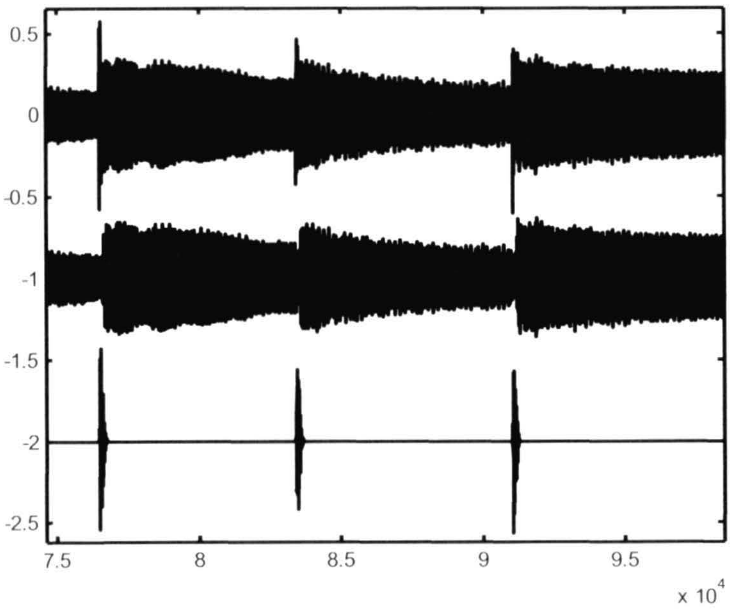


Figure 1. Example of transient synthesis from parameters (bottom) and the signal after transient subtraction (middle) from the original signal (top).

B. *Multi-resolution sinusoidal analysis*

Accurate detection of spectral peaks representing partials in the dense range of very low frequencies as well as detecting deeply modulated partials in the more sparse range of upper frequencies calls for multi-resolution spectral analysis. The proposed HSNM system employs a traditional structure of subband decomposition followed by FFT transform of different resolutions suited to particular signal properties in each subband. For practical reasons, the number of subbands is limited to three. The configuration (splitting frequencies and transform block lengths, N) is experimentally optimized by calculating the modeling error for a range of signals with all reasonable combinations of settings. Listening tests indicate, that the best configuration found in this way (table I) also offers a best subjective quality of modeling. In all resolutions, consecutive analysis frames are strongly overlapping and advanced by 6 or 12ms.

TABLE I. MULTI-RESOLUTION ANALYSIS SETUP

Subband	*Frequency range*	*Subsampling*	*FFT length, N*
1	20Hz – 310Hz	64:1	16384 (256)
2	310Hz – 2480Hz	8:1	2048 (256)
3	2480Hz – 20kHz	–	1024

Transient detection and removal before sinusoidal analysis effectively reduces pre-echo in case of impulse-like transients, however there is still a possibility that high window overlapping yields pre-echo in case of step-like transients. Therefore, a special pre-processing is applied for analysis windows marked as containing a transient in their second half. In such a case, the sequence of samples starting from detected transient position is replaced by a predictor output based on previous samples (fig. 2). A high-order (e.g. p=500) auto-regressive (LP) model is trained on data samples preceding the transient, and a sequence of zeros is passed to the input of the LP predictor while preserving its inner state after processing those original samples. The prediction signal partially replaces the original signal. This allows to avoid detection of new sinusoidal partials related to transient in frames preceding the actual transient position.

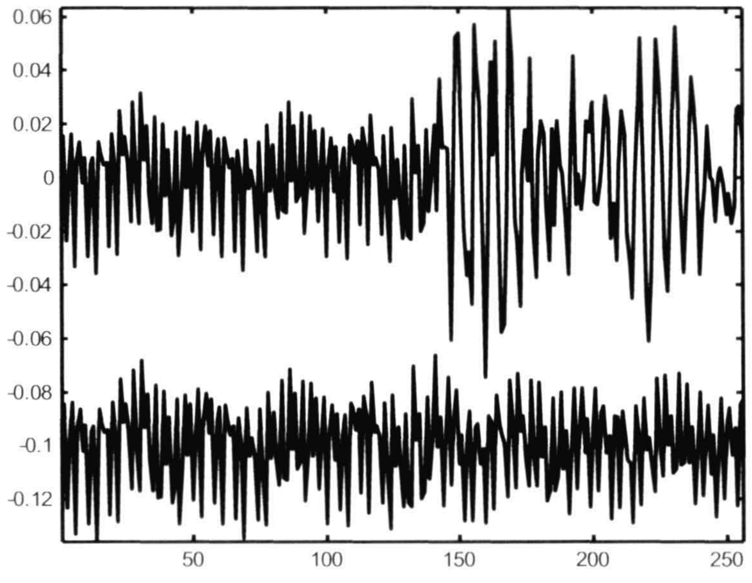


Figure 2. Example of step-like transient removal. In the case of a transient located in the second half of the analysis window, the original signal (top) is replaced by the output of a high order predictor (bottom).

Sinusoidal partials are detected by applying the standard peak picking procedure to the magnitude spectrum in each frame. Optional peak selection may be performed in order to avoid estimation and further tracking of partials which are inaudible. Fort this purpose, all peaks falling below the frequency dependent absolute threshold of hearing are rejected. Furthermore, a small clearance zone (e.g. 10Hz) is defined around each detected peak. Peaks of magnitude lower than 10dB w.r.t the maximum peak within this zone are rejected as well.

Estimation of partial frequencies, amplitudes and phase is based on the ML method with quadratically interpolated Fourier transform and takes into account the shape distortion of spectral main lobes related to frequency and amplitude modulations [11]. However, the proposed HSNM system is not fixed to any particular estimation method, and other methods may be used as well.

C. Adaptive tracking

The tracking algorithm is a result of an extensive development. It applies a carefully chosen set of criteria for finding track continuations in a collection of spectral data. The most important technique employed here is an adaptive prediction which is much more successful in tracking of modulated sounds than any simple technique taking into account only the absolute frequency difference of partials. An LP predictor is capable of learning the character of typical vibrato and tremolo from the beginning of the note and accurately predicts its further evolution [12]. Its application is motivated by the observation that pitch and intensity variations in many natural sounds are related to the motion (rocking or swinging) of player's hand which in turn may be characterized by certain mechanical resonant modes.

For each trajectory defined by a sequence of parameters $\{f_i, A_i\}$, the continuation is calculated from its existing evolution with a standard LP prediction equation,

$$\hat{f}_m = \sum_{i=1}^{P} a_i^{(f)} f_{m-i} \text{ and } \hat{A}_m = \sum_{i=1}^{P} a_i^{(A)} A_{m-i} \cdot \tag{3}$$

For all data points available in the current frame, a degree of frequency and amplitude matching λ is calculated by

$$\lambda_f \left\{\hat{f}_m, f_m\right\} = \min\left\{1 - \frac{\left|\hat{f}_m - f_m\right|}{\Delta_{\max} f_m}, 0\right\}, \tag{4}$$

where $\Delta_{\max} f_m = \max\left\{\delta_f f_m, \quad \Delta_f\right\} \quad [\mathrm{Hz}]$, and

$$\lambda_A\left(\hat{A}_m, A_m\right) = \min\left\{1 - \frac{\left|\hat{A}_m - A_m\right|}{\Delta_{\max}\left(\hat{A}_m, A_m\right)}, 0\right\} \tag{5}$$

where $\Delta_{\max}\left(\hat{A}_m, A_m\right) = \begin{cases} \Delta^{+} A & \hat{A}_m \geq A_m \\ \Delta^{-} A & \hat{A}_m < A_m \end{cases} \quad [\mathrm{dB}] \cdot$

The above measures are normalized in the range of <0,1>, and related to predefined thresholds (Δf, δf, $\Delta^{+}A$, $\Delta^{-}A$) that allow to control the sensitivity of the algorithm.

Note, that for the maximum change of frequency $\Delta_{\max} f$, both absolute difference limit (Δf) and relative difference limit (δf) is considered (set approximately to 30Hz and 3%, respectively). This is a crucial modification w.r.t. the original algorithm [2], and allows to properly cope with frequency modulation depth increasing for high-order partials of a sound spectrum, while taking into account the typical accuracy limitations of frequency estimation which is a part of sinusoidal analysis. On the other hand, for the maximum amplitude change $\Delta_{\max} A$, separate limits are defined for amplitude increase ($\Delta^{+}A$) and decrease ($\Delta^{-}A$), typically in the range of 6 to 20 dB. The joint degree of matching is calculated as $\lambda = (\lambda_f \lambda_A)^{1/2}$.

Connections are made according to a greedy rule, i.e. best matching pairs are connected in the first order, and a connection is forbidden for λ=0.

Other track continuation methods may be used optionally, such as the first order, non-adaptive prediction, where $a_i=0$ for $i>1$, or a linear trend-based prediction in log scale of frequency and amplitude. In such a case, the algorithm resorts to alternative criterion when the basic criterion does not find a matching data point. The last resort is generating zombie points that help to bridge connections over a number of frames with missing data. A sequence of limited number (e.g. 2 or 3) of successive zombie points is allowed in each trajectory by simply using the predictor outputs for f_m, and A_m, respectively.

D. Merging of trajectories

The tracking algorithm creates trajectories progressively, from previous frame to the current frame, in the direction of time. This strategy may result in missing connections, due to bad initialization of the predictor (3). Furthermore, in certain conditions, a sequence of alternating values representing a modulated partial yields a creation of several parallel trajectories with zombie points instead of one evolving according to the modulations. An additional post-processing of trajectories is aimed at increasing the continuity by merging fragmented trajectories as well as absorbing weak trajectories by a close strong neighboring one.

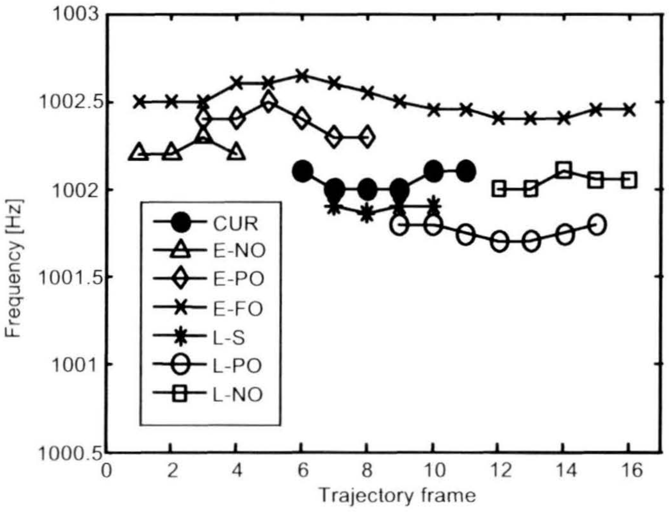


Figure 3. The classes of neighboring trajectories.

The iterative trajectory merging algorithm analyses all frames in a sequence. A set of trajectories which end in current frame is determined and sorted according to the energy of corresponding sinusoids. For every currently considered trajectory (CUR) a list of merging candidates is created. All trajectories within a small time and frequency neighborhood of CUR are assigned a specific class. Six classes are defined (cf Figure 3.): earlier–non–overlapping (E-NO), earlier–partially–overlapping (E-PO), earlier–fully–overlapping (E-FO), later–shorter (L-S), later–partially–overlapping (L-PO) and later–non–overlapping (L-NO). The best possible candidate for merging is determined in next step. All L-S candidates are discarded at the beginning, as they are redundant in current iteration. The choice between the rest of candidates is based on a degree of matching, which is essentially the same as defined in (4-5), albeit with more conservative limits of Δf and δf(10Hz and 1%, respectively). For non–overlapping candidates, an LP based extrapolation of trajectories is calculated for a number of frames in order to determine the degree of matching on a longer distance.

The actual process of merging varies depending on whether the accepted candidate is overlapping or non-overlapping. For overlapping cases, all overlapping knots are to be combined and their parameters need to be recalculated. The new values of amplitudes and frequencies are determined as

$$\hat{A} = \left|\hat{X}\right|, \quad \hat{\varphi} = \arg\{\hat{X}_k\}, \quad \hat{f} = \frac{A_k^2 f_k + A_m^2 . f_m}{A_k^2 + A_m^2}, \qquad (6)$$

where $\hat{X} = |X_k + X_m| \exp\left(j \arg\{X_k + X_m\}\right)$, $X_k = A_k \exp(j\varphi_k)$ and $X_m = A_m \exp(j\varphi_m)$ are complex representations of corresponding trajectory knots. In case of non-overlapping candidate choice, missing knots are obtained by linear interpolation of amplitude and cubic spline interpolation of frequency values. An example effect of trajectory merging is shown in Figure 4. It may be noted that this process results in combining a sequence of broken segments into a long continuous trajectory.

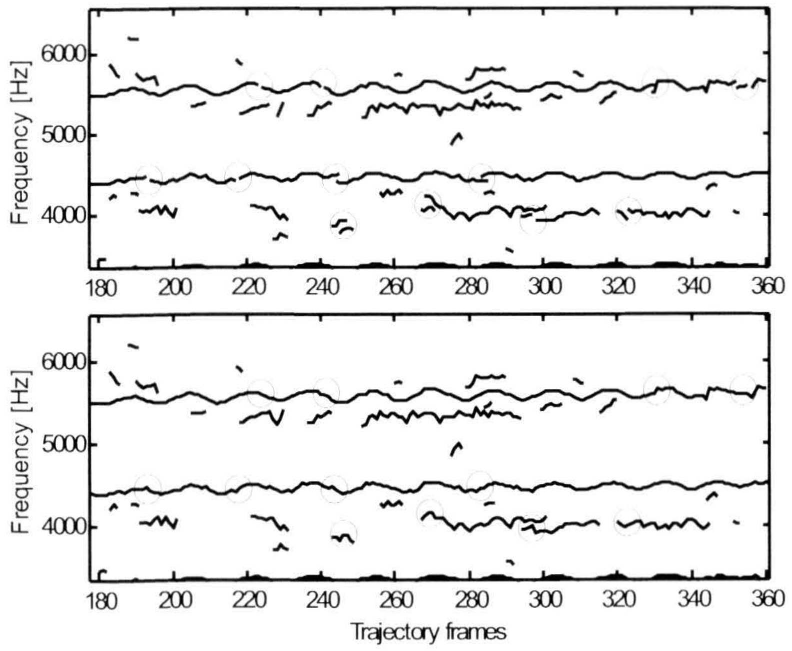


Figure 4. Trajectories before (top) and after merging (bottom). Thin-line circles indicate places where the merging was performed.

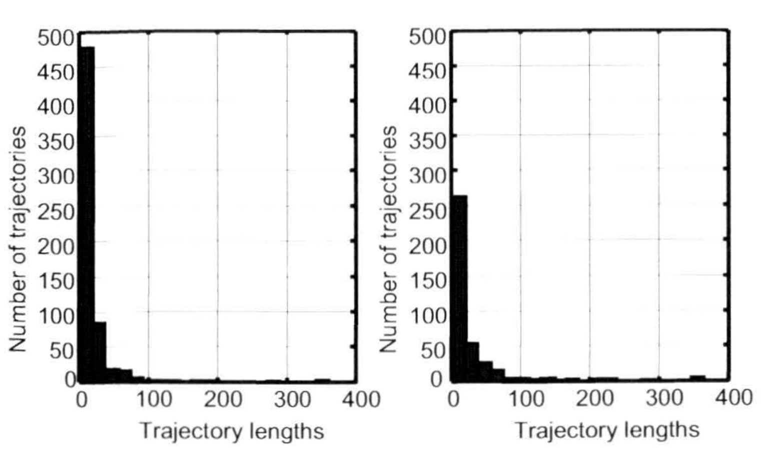


Figure 5. Example histograms of trajectory lengths before merging (left) and after merging (right).

Fig. 5 shows that after trajectory merging the number of short trajectories is significantly reduced, while the number of longer trajectories is slightly increased. The total number of trajectories after merging is usually reduced by 40%-60%. The advantage of merging is not only the reduced complexity of the model, but also reduced number of discontinuities which are responsible for potential synthesis artifacts.

E. Parameter re-estimation

The additional stage of parameter estimation aims at correcting estimation errors, potential artifacts resulting from merging, as well as estimating zombie data points. Having the trajectories already established, it is possible to estimate the amplitudes and frequencies of time-varying sinusoids yielding minimum energy of the residual. The re-estimation is performed frame by frame. In every data segment centered at current frame, a least-squares solution to a matrix equation $\underline{x}_+ = \mathbf{W}\, \underline{c}$ is computed, where

$$\mathbf{W} = \begin{bmatrix} A_1(1)\exp[j\varphi_1(1)] & \cdots & A_K(1)\exp[j\varphi_K(1)] \\ \vdots & \ddots & \vdots \\ A_1(N)\exp[j\varphi_1(N)] & \cdots & A_K(N)\exp[j\varphi_K(N)] \end{bmatrix} \qquad (7)$$

is a matrix of interpolated trajectory samples, $\underline{c}$ is a vector of complex correcting coefficients for all trajectories in the current frame, $\underline{x}_+$ is a column vector of samples of an analytic signal $x_+(t) = x(t) + \mathrm{j}\,\mathscr{H}\{x(t)\}$, and $\mathscr{H}\{\}$ is the Hilbert transform.

The elements of $\underline{c}$ obtained by solving the above equation are used to correct partial parameters by substituting

$$A_k \leftarrow A_k\,|c_k|, \text{ and } \varphi_k \leftarrow \varphi_k + \arg\{c_k\}. \qquad (8)$$

The choice of the length of segment N is quite important as it affects the accuracy of the technique in time and frequency. It is particularly essential in the low frequency range, where the segment should be long enough to accommodate many cycles of the estimated waveform. For best results, the set of trajectories is divided according to their mean frequency, and different segment lengths are used in each subset. Analysis settings shown in table 1 proved in an extensive series of experiments to deliver the most accurate results.

F. Noise modeling

The noise model is based on the frequency warped LP technique [13]. It is particularly advantageous for wideband audio, since a proper selection of the warping coefficient allows to achieve a much more accurate estimation of the residual spectrum envelope in the low frequency range than the traditional auto-regressive (LP) model (cf. fig. 6).

The residual is analyzed in consecutive overlapping frames of fixed length (e.g. 10ms). The warped linear prediction coefficients are estimated in each frame as well as the energy of the prediction error. During re-synthesis, these parameters are used to generate an appropriately shaped random signal. Additional HP filter (2nd order, f_c=300Hz) and an LP filter (2nd order, f_c=12kHz) are employed for compensating the over-estimated power density in the very low and very high frequencies, as shown in fig. 6.

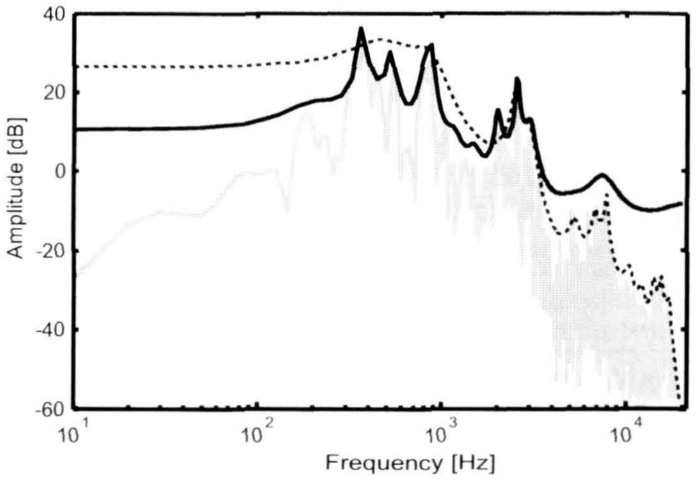


Figure 6. Spectral envelope modeling of the original signal (grey) by a standard LP model (dashed line) and the WLP model (black) of the same order (30). Note the increased accuracy in low frequency range in WLP versus an unnecessarily high accuracy in the high frequency range in LP.

G. Signal synthesis from the model

The synthesis procedure is quite straightforward and may be performed in any order. Sinusoidal partials are synthesized sample by sample from the amplitude, frequency and phase data after linearly interpolating the amplitude in the log (dB) scale, and interpolating the phase with a cubic spline polynomial [2]. For pitch or speed transformations, phase data needs to be recalculated based on the integral od instantaneous frequency. Noise and transient components are synthesized from respective data sets and mixed with the sinusoidal part.

IV. Reconstruction Quality

The HSNM system described in this paper has been thoroughly tested and the reconstruction quality has been assessed in many experiments. For the purpose of this paper, a blind listening test has been organized according to the MUSHRA methodology [14]. Twelve subjects participated in the test, evaluating in a continuous subjective scale the anonymous reconstructed signal against known as well as a hidden reference (the original), for a collection of music excerpts representing various modeling challenges: a solo piano, solo harpsichord, a violin+acoustic guitar, a vocal quartet, a dynamic pop and RnB music, a choral piece, and a symphonic orchestra piece. The results (cf. fig. 7) indicate, that in many cases the listeners could not reliably identify the re-synthesized signal.

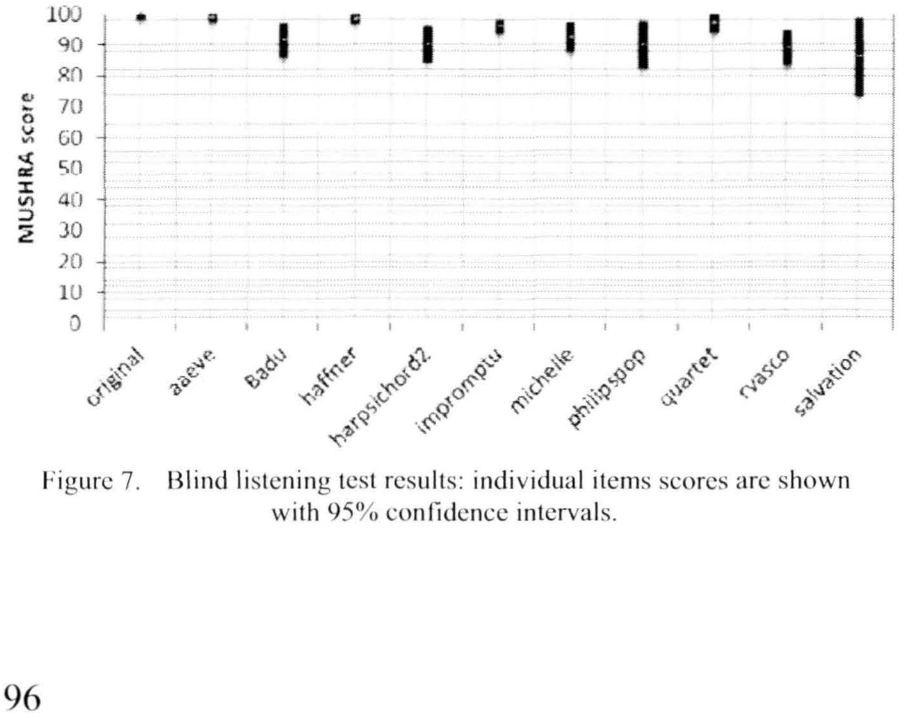


Figure 7. Blind listening test results: individual items scores are shown with 95% confidence intervals.

The reader may also individually asses the quality of reconstructed signals by visiting the project homepage at http://www.multimedia.edu.pl/audio_research/.

V. Conclusions

A hybrid sinusoidal modeling system that offers a near transparent audio quality even for complex and dynamic music has been described in the paper. This high reconstruction fidelity is achieved thanks to introduction of a dedicated transient model, as well as numerous enhancements within the traditional sinusoidal modeling scheme. The applications of this system include high quality parametric audio coding, source separation, pitch and time scale transformations, and other special effects.

Acknowledgment

The work has been supported by publish founds for scientific research.

References

[1] J. Beauchamp, "Analysis and Synthesis of Musical Instrument Sounds," in Analysis, Synthesis, and Perception of Musical Sounds, J. Beauchamp, Ed. Urbana: Springer, 2007

[2] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Trans. Acous., Speech, Sig. Proc., vol. 34, no. 4, Aug. 1986

[3] X. Serra and J.S.Smith III, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition," Computer Music Journal, vol. 14, no. 4, 1990

[4] M. Goodwin, "Residual modeling in music analysis-synthesis," Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP-96, 7-10 May 1996

[5] X. Rodet, "Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models," IEEE Time–Frequency and Time–Scale Workshop, Coventry, UK, 1997

[6] R. Badeau, R. Boyer and B. David, "EDS Parametric Modeling and Tracking of Audio Signals", Int. Conf. on Digital Audio Effects, DAFx'02, September 2002

[7] G. Peeters, X. Rodet, Signal Characterization in terms of Sinusoidal and Non-Sinusoidal Component, Proc. 1st Digital Audio Effects Conference (DAFx'98), Barcelona, 1998

[8] G. Peeters, X. Rodet, SINOLA: A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum, in Proc. Int. Computer Music Conf., Beijing, October 1999

[9] C. Duxbury, J. P. Bello, M. Davies and M. Sandler, "Complex Domain Onset Detection for Musical Signals," 6th Int. Conference on Digital Audio Effects DAFx'03, London, UK, 2003

[10] A.C . den Brinker, E.G.P . Schuijers, and A.W.J . Oomen, "Parametric Coding for High-Quality Audio", 112th Convention of the Audio Engineering Society, Munich, 2002

[11] M. Abe, J. O. Smith III, "Design Criteria for Simple Sinusoidal Parameter Estimation Based on Quadratic Interpolation of FFT Magnitude Peaks," Proc. 117th Convention of the Audio Engineering Society, San Francisco, Oct. 2004

[12] M. Lagrange, S. Marchand, M. Raspaud, J. B. Rault, "Enhanced Partial Tracking Using Linear Prediction," Int. Conf. on Digital Audio Effects, DAFx'03, London, Sept. 2003

[13] A. Härmä, et al, "Frequency-Warped Signal Processing for Audio Applications", J. Audio Eng. Soc., vol. 48, no.11, 2000

[14] ITU-R, "Method for the subjective assessment of intermediate quality level of coding systems", ITU-R, Tech. Rep. BS. 1534-1, Rec. 2003

Detection of Tampering in Lossy Compressed Digital Audio Recordings

Rafał Korycki
Institute of Radioelectronics
Warsaw University of Technology
Warsaw, Poland
r.korycki@ire.pw.edu.pl

***Abstract* — This paper addresses the problem of tampering detection and discusses one of the methods used for authenticity analysis of digital audio recordings. Presented approach is based on checking frame offsets in audio files compressed by using perceptual audio coding. This method can be further improved by applying histogram analysis of modified discrete cosine transform spectrum and detection of maximum number of nonzero spectral coefficients. The influence of compression algorithms' parameters on detection of forgeries are presented by applying AAC and Ogg Vorbis encoders as examples. The effectiveness of tampering detection algorithms proposed in this paper is tested on a predefined music database and compared graphically using ROC-like curves.**

***Keywords* — component tampering detection, digital audio authenticity, lossy compression, frame offsets, MDCT, spectral coefficients**

I. Introduction

Digital recording authenticity analysis is extremely important in juridical proceedings. Recordings may be useless as evidence when there is no proof that they are original and they have not been tampered. It is difficult to detect traces of montage in the era of digital audio files, due to the fact that currently available technologies e.g. free sound editing software allow a forger to change the meaning of uttered sentences without audible artifacts. Therefore, any tool that helps to evaluate digital audio authenticity may be of great importance to forensic audio experts. Nowadays, the most accurate and commonly used authentication method is based on electric network frequency (ENF) criterion [5]. This approach utilizes random frequency fluctuations of the mains signal emitted by the electric network which are inadvertently induced in electronic circuits of recording devices. Therefore, its effectiveness is strictly dependent on the presence of mains signal in the recording, which occurs rarely. Furthermore some of the aspects of this approach, such as accurate measurements of ENF as well as searching and comparing the fluctuation patterns with a reference database are still being subjects of a scientific debate [4], [5], [10].

Recently, much attention is paid to authenticity analysis of compressed multimedia files . Several solutions were proposed for detection of double quantization in digital video which may result from multiple MPEG compression as well as from combining two videos of different qualities [2]. Blocking periodicity analysis in JPEG compressed images was also investigated according to differences in quantization errors between neighboring blocks [1]. Due to differences between audio and picture compression algorithms it is impossible to adapt these methods in audio authenticity analysis. Instead, the detection of forgeries in compressed audio recordings must be based on other mathematical properties. Grigoras described statistical tools used to detect traces of audio recompression data to assess compression generation and also to discriminate between different audio compression algorithms [6]. Liu et al. presented novel approach to detect double MP3 compression by extracting the statistics of the modified discrete cosine transform (MDCT) spectral coefficients of MP3 signals, followed by applying a support vector machine [12]. Moreover, Huang et. al presented the numbers of small-value MDCT coefficients as features to discriminate fake-quality MP3 from normal MP3 [9] as well as to perform authenticity analysis [8].

The remainder of this paper is organized as follows. In Section II a brief introduction to lossy compression is given including short explanation of MDCT properties. Analysis of spectral components and discussion on frame offsets is presented in Section III. Shown in Section IV are three tampering detection algorithms proposed by the author, which effectiveness in detection of forgeries will be proven based on database of music tampered recordings. Obtained results are discussed in section V.

II. Lossy Compression

Analyzing lossy compression algorithms, it is worth taking into consideration the principles that are relevant to the authentication process, i.e. spectral decomposition and quantization. The generic structure of perceptual audio coder consists of: (i) filterbanks and transform blocks where the input samples are converted into a subsampled spectral representation, (ii) perceptual model in which the signal's time-dependent masking threshold is estimated, (iii) quantization block where the quantization noise is masked, (iv) coding block, in which relevant information are packed into a bit stream [7].

In MPEG Layer 3 (MP3) encoder a sequence of 1152 input samples are poliphase filtered into 32 frequency subbands. Subsequently, Modified Discrete Cosine Transform (MDCT) is applied to the time frames of subband samples and each of the 32 subbands is split afresh into 18 subbands creating a granule with a total of 576 frequency coefficients. To reduce artifacts caused by time-limited operation on the signal, the windowing is applied. Depending on the degree of stationarity, the psychoacoustic model determines, which of four types of window is used [8], [10]. The Advanced Audio Coding (AAC) and Ogg Vorbis algorithms employ only MDCT filterbank which is switched between resolutions of 1024 and 128 spectral lines, depending on the stationary or transient character of the input signal. Moreover, in AAC the shape of the transform window can be adaptively selected between a sine window and a Kaiser-Bessel-derived (KBD) window. Afterwards, the spectral coefficients are quantized, given the masking thresholds estimated by the psychoacoustic model and then coded using a set of Huffman code tables [7], [10].

The MDCT is central for tampering detection algorithms, due to its properties. *N*-sized MDCT is computed based on 2*N* samples taken with 50% overlapping. Applying Inverse-MDCT transform to the frame of spectral coefficients yields therefore 2*N* time-aliased samples. These distortions are canceled by the overlap-and-add (OLA) procedure, which consists in computing an Inverse-MDCT, based on the previous and the next frame, multiplying each of the aliased segments by its corresponding window function and summing up overlapping time segments. For signals with local symmetry, the MDCT coefficients are frequently reduced to zero [8], [11].

III. Analysis of MDCT Spectral Components

While encoding process is applied, spectral coefficients are quantized and some of them are assigned a zero value. This occurs for masked components as well as for unmasked parts, due to the probability distribution of the spectral coefficients and the compression ratio [7]. Within the decoded signal, the troughs in a logarithmic spectral representation are visible only if identical framing offset to the one used for encoding is applied [8]. It is essential to employ a correct decomposition algorithm, including proper window length and shape as well as the same filterbank type as used during the encoding process. This is because encoders' specifications usually do not define the exact steps for processing input data. The algorithms can therefore function quite differently and still satisfy the standard [7].

Shown in Fig. 1 are MDCT coefficients of a decoded audio recording computed by using analysis window with one sample left shift (offset=-1), no sample shift (offset=0) and one sample right shift (offset=+1) from the encoder frame grid, respectively. To process data and perform lossy compression AAC algorithm was used. As may be inferred from these examples, even a shift by one sample is sufficient to conceal the presence of characteristic zero values in the spectral representation of analyzed signal [7]. The troughs are visible only if the same frame offset is chosen as was employed during the encoding process.

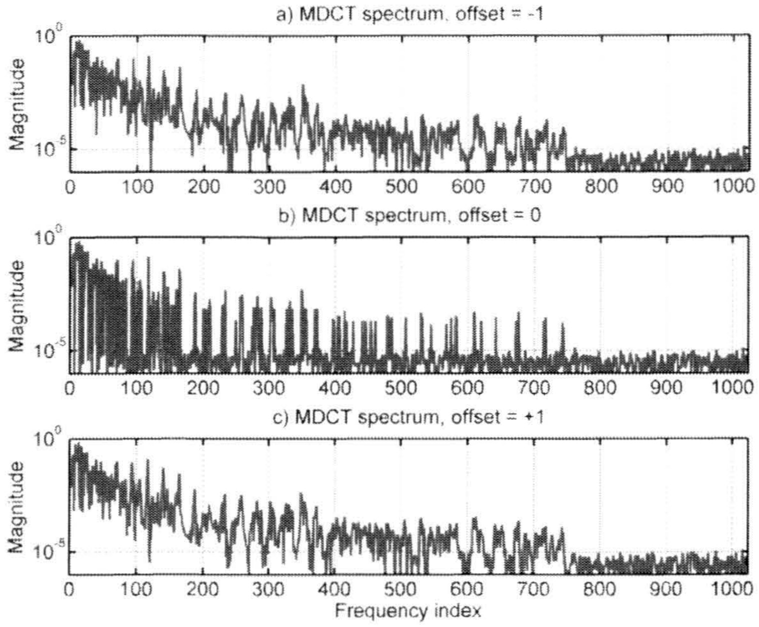


Figure 1. MDCT coefficients of a decoded audio recording computed by using analysis window with: a) one sample left shift (offset=-1), b) no sample shift (offset=0), c) one sample right shift (offset=+1) from the encoder frame grid, respectively. The magnitude is shown in the logarithmic scale.

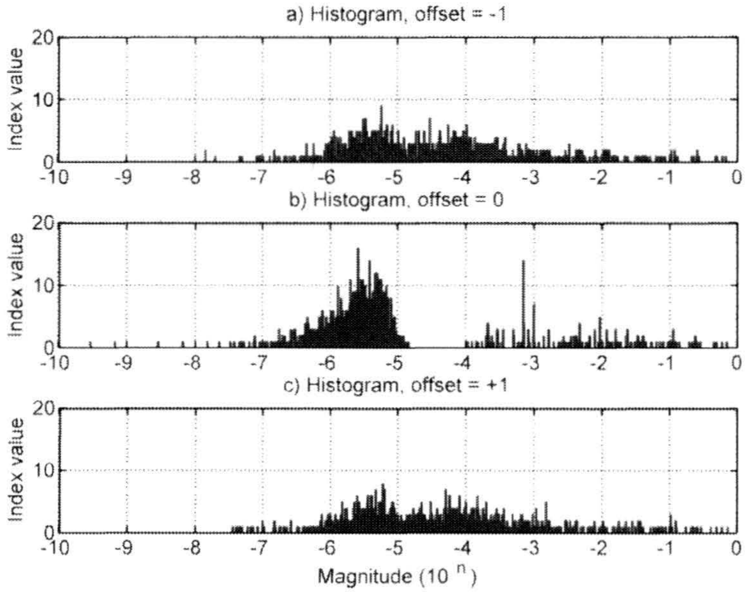


Figure 2. Histogram of MDCT spectrum of a decoded audio recording computed by using analysis window with: a) one sample left shift (offset=-1), b) no sample shift (offset=0), c) one sample right shift (offset=+1) from the encoder frame grid, respectively

The MDCT coefficients yield apparent significant distance, between the lowest peak and the highest valley of the spectrum for frame offset equal to zero. Shown in Fig. 2 are MDCT spectrum histograms of a decoded audio recording computed with one sample left shift (offset=-1), no sample shift (offset=0) and one sample right shift (offset=+1) from the encoder frame grid, respectively. As can be seen, spectral components are not present between 10^{-5} and 10^{-4} value, if the proper frame offset is chosen. It means that the magnitude of none of the spectral lines appears in this range of values. This phenomenon will be further utilized in automatic forgery detection algorithms.

IV. Tampering Detection Algorithms

The effects described in Chapter III are utilized in detection of forgeries in lossy compressed digital audio recordings. A number of active spectral coefficients (NAC) is calculated as proposed in [8] with one sample step offset:

$$M_k^{(j)} = \log_{10}\left(\max\left(X_k^{(j)} \cdot X_k^{(j)} \cdot 10^{10}, 1\right)\right), \qquad (1)$$

where j is frame offset and k is frame index. Signal samples are multiplied by sliding window with predefined length and shape prior to the MDCT. Additionally, computations are performed for each of four types of windows used by the encoder, which enables the algorithm to analyze real audio recordings. When current value of a frame shift equals multiplication of a window length, NAC reaches its minimum. Hence, if recording is tampered, the offset between adjacent minimums of NAC function outside the forgery position is other than current window length.

Fig. 3 illustrates a fragment of audio recording compressed and decoded by AAC algorithm (with 128 kbps constant bit rate and 44,1 kHz sampling frequency) in which four edits were made. NAC function related to the frame offset was computed for each of four window types. To simplify the analysis, KBD window for long blocks was not applied in the AAC algorithm. Application of a short window requires a window switching sequence which is also apparent in Fig. 3. Another crucial observation is that edit points are accompanied by maximums of NAC function. Despite the fact that they may be found in other places where forgeries do not occur, these extreme values are utilized to improve robustness of tampering detection algorithm.

As stated above, the NAC minimums occur only for frame shifts equal to multiplications of applied window length. Any modification of audio file, including cutting off or pasting a part of audio recording causes a disturbance within this regularity. Based on the described phenomenon, three algorithms were developed. The ALG 1 consists in analysis of differences between adjacent minimum positions of NAC function. Depending on a window used during the encoding process, measured distances are matched with predefined pattern. Observed variances are compared parallel to each other for every window and stored as a possible indication of tampering.

The ALG 2 algorithm employs additional histogram analysis. For each minimum of NAC function the absence of spectral components of magnitude values between 10^{-5} and 10^{-4} is examined. These two results are logically multiplied and the outcomes are treated in the same manner as minimums of NAC in ALG 1. In the last algorithm (ALG 3), detection of maximum values of NAC function is employed. The results are then maximized within the area of window length and multiplied by the outcomes from ALG 2. The applied solution radically minimizes a number of false detections of forgeries.

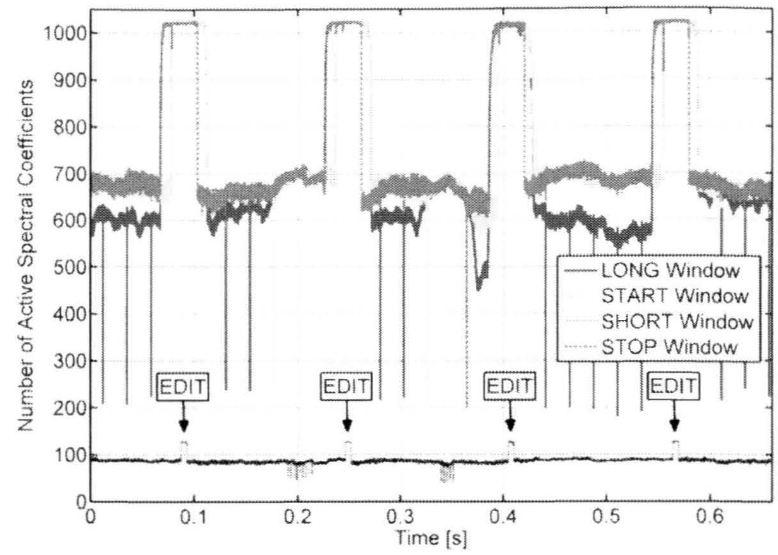


Figure 3. NAC function related to the frame shift computed for a fragment of audio recording in which four edits were made. Audio recording was compressed and decoded using AAC algorithm with 128 kbps constant bit rate and 44,1 kHz sampling frequency

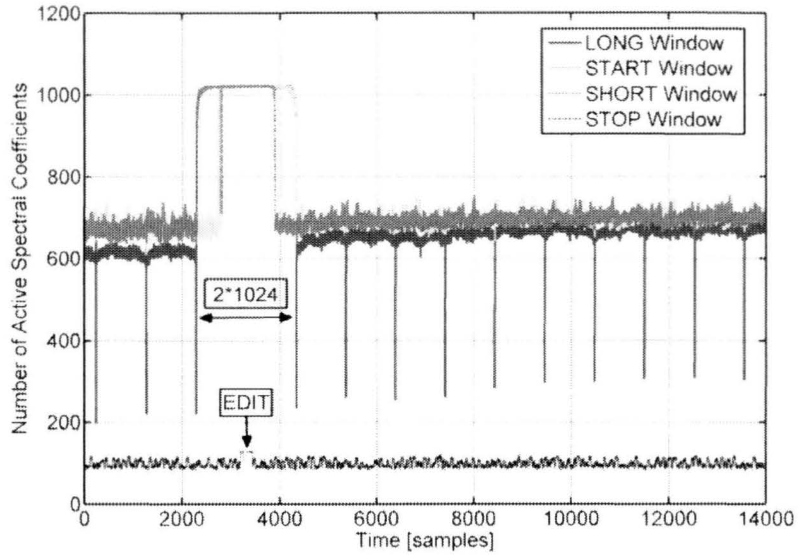


Figure 4. NAC function related to the frame shift computed for a fragment of audio recording in which one edit was made. A part of audio file lasting the exact multiples of applied window length was removed

Theoretically, cutting out a part of audio file lasting the exact multiples of applied window might not have caused detectable disturbances in frame offset, therefore analyzed recording might be recognized as unaltered. However, application of MDCT computed for blocks of signal samples taken with 50% overlapping still allows to find the trace of this forgery. Shown in Fig. 4 is a fragment of an audio file compressed and decoded by AAC algorithm (with 128 kbps constant bit rate and 44,1 kHz sampling frequency) from which a fragment of the recording lasting about 23 ms (10 · 1024 samples) was removed. As can be seen the distance between adjacent minimum positions of NAC function is other than 1024 for long blocks and the value of NAC function reaches its maximum in the vicinity of edit point.

V. Results and Discussion

Described algorithms are examined for their usefulness as tools for detection of forgeries in lossy compressed audio recordings. The tests were conducted on a music database consisting of 15 music tracks with harmonic components and slowly changing audio background. The recordings were compressed and decoded using two different encoders with three different bit rates. In each of these tracks, 30 second long fragments were selected for further processing. The music database was prepared to mimic real forensic recordings, usually made in a noisy environment. In all of sampled fragments 21 deletions were performed at randomly selected locations and of randomly selected durations, however no longer than one second. Therefore, 315 forgeries were produced and subjected to further examinations.

The algorithms described in Section IV were employed to detect edits performed in recordings from the music database. Thereafter, true acceptance ratios (TAR) and numbers of false acceptances (FA) were computed, given changing detection thresholds of minimums of NAC function. The TARs are obtained based on the number of detected edit points, divided by the known total number of forgeries. Shown in Table I and Table II are computational results obtained for two different encoders: AAC and Ogg Vorbis, respectively. Figures 5-6 show modified receiver operating characteristics (ROC) plotted for each of employed algorithms, encoders and bit rate values. Typically, ROC curve illustrates the performance of a binary classification system when its discrimination threshold is being modified. It is created by plotting the true positives out of the total number of positives as compared to the false positives out of the total number of negatives. The ROC-like curves used for the purpose of this article (Figures 5-6) depict true acceptance ratios in the function of the number of false acceptances, given changing detection threshold of minimums of NAC function.

TABLE I. Simulation Results for Edited Fragments from the Music Database compressed and decoded using AAC encoder

Bit rate [kbps]	ALG 1		ALG 2		ALG 3	
	TAR	*FA*	*TAR*	*FA*	*TAR*	*FA*
64	1,0000	1325	1,0000	1017	0,9905	6
96	1,0000	324	1,0000	140	0,9968	0
128	1,0000	214	1,0000	81	0,9810	0

TABLE II. Simulation Results for Edited Fragments from the Music Database compressed and decoded using Ogg Vorbis encoder

Bit rate [kbps]	ALG 1		ALG 2		ALG 3	
	TAR	*FA*	*TAR*	*FA*	*TAR*	*FA*
64	1,0000	26	1,0000	21	1,0000	0
96	1,0000	48	1,0000	36	1,0000	0
128	1,0000	31	1,0000	22	0,9873	0

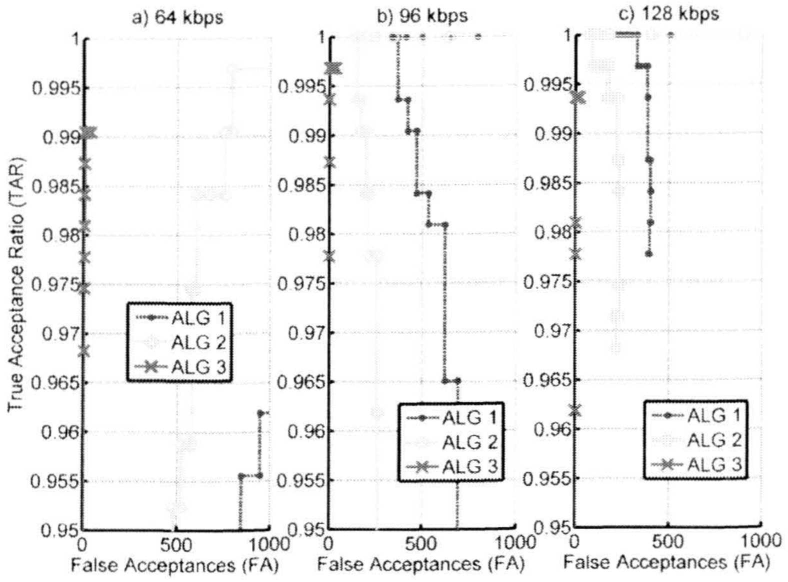


Figure 5. ROC-like curves obtained during testing procedure of proposed algorithms executed on edited fragments from the music database. AAC compression algorithm was applied

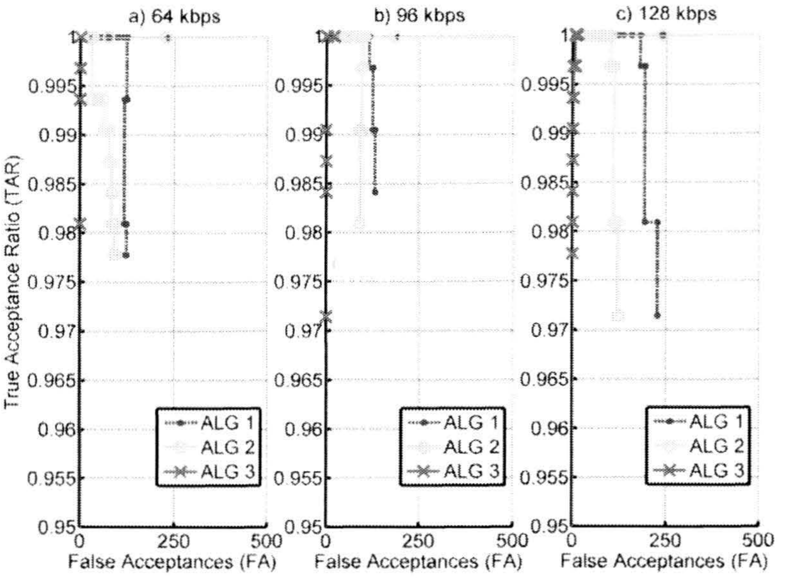


Figure 6. ROC-like curves obtained during testing procedure of proposed algorithms executed on edited fragments from the music database. Ogg Vorbis compression algorithm was applied

As may be seen, ALG 1 and ALG 2 algorithms, which are based on detecting minimums of NAC function give superior forgeries detection ratio with intolerable number of false acceptances. These results are coincident with the outcomes presented in [8] and obtained for MP3 encoder using probably the long blocks only. In contrast, the ALG 3 algorithm yields slightly lower TAR values, while the number of false acceptances is still reduced almost to zero. The results obtained for AAC encoder at 64 kbps (Figure 5a) are significantly affected compared to all other bit rates, i.e. lower TAR values and higher number of false acceptances. This phenomenon is caused by default bandwidth limitation (to about 6 kHz) applied during encoding with the given bit rate.

VI. Summary

The presented approach to analysis of nonzero spectral coefficients obtained based on MDCT transform can be successfully applied to those recordings in which the alleged forgery was made *after* audio files were decoded to lossless format and which were not encoded afresh. The methods proposed in the literature are improved by employing analysis of each of the four types of the transform window. Furthermore, histograms of MDCT spectra and detection of maximum number of nonzero spectral coefficients are applied to enhance robustness of initial algorithm.

Presented methods consisting in detection of frame offset of the compressed audio files can be successfully applied by forensic experts to detect forgeries in lossy compressed digital audio recordings. The value of tampering detection ratio for the given number of false acceptances equal to zero, is greater than 0.98. This allows the method to be recognized as a robust assistance in authenticity investigation process. Notwithstanding its robustness and accuracy, this approach requires more thorough research, especially in case of algorithmic differences between particular types of encoders.

REFERENCES

[1] Y. Chen and C. Hsu, "Image tampering detection by blocking periodicity analysis in JPEG compressed images", in *Proc. IEEE 10th Workshop on Multimedia Signal Processing*, Cairns, Queensland, Oct. 8-10, 2008, pp. 803-808.

[2] H. Farid and W. Wang, "Exposing Digital Forgeries in Video by Detecting Double Quantization" in *Proc. ACM Multimedia and Security Workshop*, Princeton, 2009.

[3] R. Geiger, J. Herre and S. Moehrs, "Analysing decompressed audio with the Inverse Decoder - towards an operative algorithm", in *Proc. 112th AES Convention*, Munich, Germany, May 10-13, 2002.

[4] Z. Geradts, M. Huijbregtse, "Using the ENF criterion for determining the time of recording of short digital audio recordings", *IWCF Proceedings of the 3rd International Worshop on Computational Forensics"*, 2009

[5] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis", *Forensic Science International*, vol. 167, pp. 136-145, 2007.

[6] C. Grigoras, "Statistical Tools for Multimedia Forensics: Compression Effects Analysis", in *Proc. 39th AES International Conference Audio Forensics*, Hillerod, Denmark, June 17-19, 2010.

[7] J. Herre, M. Schug, "Analysis of Decompressed Audio – The Inverse Decoder", *Proceedings of the 109th AES Convention*, Los Angeles, USA, Sep. 22-25, 2000.

[8] J. Huang, Z. Qu and R. Yang, "Detecting digital audio forgeries by checking frame offsets", in *Proc. of the 10th ACM workshop on Multimedia and security table of contents, International Multimedia Conference*, Oxford, 2008, pp. 21-26.

[9] J. Huang, Y.Q. Shi, R. Yang, "Defeating Fake-Quality MP3", *Proceedings of the 11th ACM workshop on Multimedia and security*, Sep. 7-8, 2009, Princeton NJ.

[10] R. Korycki, "Methods of Time-Frequency Analysis in Authentication of Digital Audio Recordings", *International Journal of Electronics and Telecommunications*, vol. 56, no. 3, pp. 257-262, 2010.

[11] R. Korycki, "Research on Authentication of Compressed Audio Recordings, in *Proc. 14th International Symposium on Sound Engineering and Tonmeistering ISSET 2011*, Wroclaw, Poland, May 19-21, 2011.

[12] Q. Liu, M. Qiao, A.H. Sung, "Detection of double MP3 compression", *Cognitive Computing*, Vol. 2 (4), pp. 291-296, 2010.

MusicEar – a System for Real Time Analysis and Archivization of Violin Sound

Ewa Łukasik, Marcel Makowski, Leszek Malchrowicz, Tomasz Nawracała, Adam Robak
Institute of Computing Science, Poznan University of Technology
Poznań, Poland

Abstract **— The paper presents a prototype of a portable PC-based system, called MusicEar, able to perform real time analysis, archiving and visualization of violin sound and thus support violin sound quality assessment. MusicEar visualizes the frequency spectrum in different scales including the Duennwald frequency regions designated specially for timbre characteristics of violins. It also calculates the instantaneous parameters based on Long Term Averaged Spectrum (LTAS) and stores them in a database. It enables comparison of characteristics of instruments sound already saved in the database.**

Keywords: violin, musical instruments, sound analysis, sound quality assessment, real time system, audio database

I. Introduction

The violin is a string instrument, with four strings tuned in perfect fifths. It is known as "one of the most astounding and complicated acoustical devices ever created by and for the human nervous system. A product of the glorious musical and artistic developments of the Baroque era, it has been the subject of scientific investigation since the early 19th century." [5] Research in violin acoustics has been concerned with its mechanical subsystems such as strings, wood, soundpost, plate vibrations, body vibrations, radiation etc. Each epoch adds new violin research contexts, related to the technologies and methods developed in a given period of time. The recent advances in signal processing enabled digital analysis of violin sound and advances in machine learning and soft computing that have found application in the domain of musical instruments classification are promising also in the classification and recognition of individual violins.

Most of the research in the domain of musical instruments classification based on their sound is performed on the recorded sound samples. However there are situations, where a real time violin-driven sound analysis and immediate violin sound classification is required, as e.g. in a violinmaker workshop or acoustician's laboratory. Application of real time analysis of violin sound might be also useful during violinmaker competitions, where violin sound quality is assessed by a jury. In all these cases it would be desired to store the results and just analyzed sounds.

There exist general purpose commercial real time spectral analyzers, often used in acoustical laboratories, equipped in multiple computing modules e.g. Artemis [15] or Spectra Plus [16], which provide various acoustical parameters in real time. The goal of our team is to construct a specialized system, which would characterize violin sound on the basis of a few robust and specific features. Current computer technology enables us to design sound analyzer according to the special needs of acousticians or sound quality researchers.

As a proof of the above concept a system MusicEar was built which allows to analyze, visualize, archivize and compare parameters of violin sound in real time. The system stores all waveforms, spectra and calculated features in the special purpose database so that its content may serve as a dataset for further research. MusicEar system is a part of a long term project devoted to inducing jury's preferences in terms of acoustic features of violin sounds.

The Institute of Computing Science at Poznan University of Technology has been collaborating for several years with the Henryk Wieniawski Musical Society – the main organizer of the International Violinmakers Competition held every five years in Poznań. One of the results of this cooperation was a benchmark collection of violin sounds AMATI [10] that was the subject of intensive investigations in the Institute of Computing Science and other universities e.g. [13]. In 2011, during the 12th International Henryk Wieniawski Violinmakers Competition, the organizers and the jury agreed to perform tests of the MusicEar system on competing violins during the final stage of the competition.

The paper is structured as follows. Section 2 discusses the parameters of violin sound which are calculated by the MusicEar system, Section 3 presents the architecture and functionality of the MusicEar system, Section 4 describes its interface, Section 5 presents the tests performed during the violinmakers competition and Section 6 concludes the paper.

II. Parametric Representation of Violin Sound

A. Duenwald parameters

The study of acoustic properties of the violin has a long history [5][6] and there is still a significant interest in this area [1][2]. As it was mentioned earlier in the Introduction each epoch adds new violin research contexts, related to the technologies and methods developed in a given period of time. Considerable amount of research effort has went to the discovery the most important violin resonances and to relate them to violin quality. Hutchins [6] has found a strong relationship between the quality of an instrument and a frequency distance dividing two resonances, called A1 and B1 at frequency range 485-540Hz. This knowledge is currently used in research studies [13].

TABLE I. DUENWALD RANGES OF FREQUENCY FOR DETERMINING VIOLIN SOUND QUALITY

Ranges of frequency	Region
190Hz - 650Hz	A
650Hz – 1300Hz	B
1300Hz – 1640Hz	C
1640Hz – 2580Hz	D
2580Hz - 4200Hz	E
4200Hz – 6400Hz	F

Jansson [4,8] and Gabrielsson [4] suggested using Long Time Average Spectra, LTAS, which quickly give a "constant" frequency characteristics, resembling violin radiativity diagrams obtained during experiments with violin body excitement. Duennwald [3] has introduced frequency regions of interest for distinguishing between quality of violins. The Duennwald frequency bands are presented in Table I.

Duenwald has measured the overall loudness in these regions and proposed the following rules:

- If loudness in range B >> sum of the A,C and D regions, the sound is very *nasal,*
- If average B << average ACD, the sound is very *unnasal*,
- If average F >> average DE the sound is very *harsh*
- If average F << DE the sound is very *clear*.

Duenwald analyzed individual sounds of old Italian violins, violins of old masters, violins made by masters after 1800, factory violins and violins made by hobby makers. The quality parameters consisted of the numbers representing fractions of *unnasal* and *clear* sounds. The same regions have been recently investigated by Buen for Old Italian Violins [1,2] and proved to be useful to determine the quality of the violin.

The MusicEar system calculates the spectrum in Duenwald frequency regions, visualizes it in real time and archivizes signal power in these regions in the database. The results are now investigated by the authors and the parameters related to violin sound quality will find application in the next release of the system.

B. *Bark scale*

The *Bark scale* is a psychoacoustical scale proposed by Eberhard Zwicker in 1961 [14]. The scale ranges from 1 to 24 and corresponds to the first 24 *critical bands* of hearing, which are the rough approximates of frequency bandwidths of the auditory filters. Bark scale has been used in the first MPEG audio coding standards, as it is closely related to masking phenomena. In the audio classification and recognition tasks it is less often used than the *mel scale* (derived from the word *melody*) by Stevens, Volkman and Newman in 1937 [11], but it was introduced in MusicEar as a simplest perceptual scale to visualize frequency characteristics of a violin sound (24 bands).

C. *MPEG-7 and other timbral features*

In the last decade for the purpose of Music Information Retrieval a large collection of descriptors was used, including numerous timbral features [9][12]. Some of them are included in the MPEG 7 standard [7] and used for musical instruments classification. Usually for the classification of musical instruments all of them are taken into account with various weights yielded by soft computing methods.

In the prototype MusicEar system presented in this paper only a few parameters are calculated meaningful for violin sound. They are related to two criteria used by violinmakers jury: volume and timbre of sound. The easiest way to measure volume is calculating the total *energy* of the signal. This is the first parameter calculated, normalized by the number of samples N.

Five spectral parameters have been used to characterize the timbre. The simplest popular measure for characterizing timbre is *Spectral Centroid* – the center of gravity of the spectrum. Its high value indicates brightness of the sound. In MusicEar it is also calculated in the way proposed in the MPEG-7 standard – as *Audio Spectrum Centroid* calculated in log scale with reference to 1000 Hz.

Another feature related to timbre is *Spectrum Spread* – second central moment of the spectrum. It gives indications about how the spectrum is distributed around its centroid. The analogous descriptor in MPEG-7 is *Audio Spectrum Spread*, which is calculated in logarithmic scale with reference to 1000 Hz.

The fifth basic parameter describing spectrum shape is *Spectral Flatness*, which indicates to what extent a signal is self-correlated. This parameter is usually used to distinguish between tonal and noisy signals. In case of violin sound, which is mainly tonal, it may be a subtle indicator of noisiness of the signal, related e.g. to the easiness of violin response.

The signal is also presented as a waveform and spectrum in linear frequency scales.

III. FUNCTIONALITY AND ARCHITECTURE OF THE MUSICEAR SYSTEM

The MusicEar system has been designed for PC computer, preferably portable, to work with the external audio interface. It supports only one audio channel. MusicEar has a modular structure consisting of four modules with the following functions:

- sound acquisition,
- sound analysis,
- visualization,
- database.

A. *MusicEar functionality*

The system works in the following way. The signal is acquired with an external sound card and recorded in the database frame by frame. The instantaneous spectra in three different frequency scales and six parameters are calculated and graphically visualized on the fly. After the recording is finished, the Long Term Averaged Spectra (LTAS) are calculated for the whole recording and graphically visualized together with numerical values of the features for the averaged sound waveform. Both - instantaneous and averaged characteristics are stored in the database.

To be able to fulfill the condition of real-time analysis, only 16 bit samples with the sampling frequency 44,1 kHz were acquired despite the fact that the external sound card supports higher values of both parameters. FFT window length is 8192 samples, giving 5,38Hz resolution in frequency domain.

Instantaneous spectra and parameters are displayed in the appropriate charts to enable the user to visually analyze the sound while listening. After the recording is finished the averaged spectra and parameters values are displayed in the same chart windows giving further possibility to find the characteristic features of the sound image in the regions of interest. These averaged values, together with signal waveforms and certain textual labels entered by the user are added to the database and archivized.

The database is an important part of the system, since it archivizes where all parameters and sound samples. It is possible to browse all information stored in the database and make the comparison of data for various violins in a new window. The user may have access to the table with database content, play a desired sound, export .wav files and parameters values in .csv format, compare and delete data. Data are compared in a new window, where pairs of spectrum charts and parameters values are placed side by side.

B. *MusicEar Interface*

Fig. 1 presents the MusicEar system main window. The field marked with number 1 contains text boxes for inserting information identifying the violin.

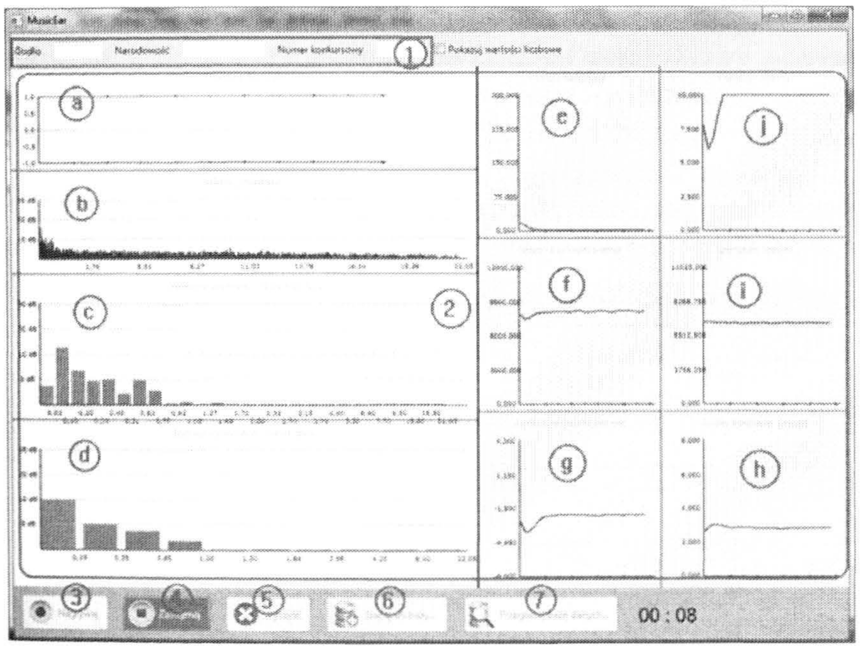


Figure 1. MusicEar main window in the instatenous mode (frame –by-frame representation): a) signal waveform, b) spectrum in linear scale, c) log spectrum in Bark scale, d), energy in Duennwald regions, e) signal energy, f) center of gravity, g – ASC, h – ASS (frame by frame), i) SS, j) SF. Function keys: 3) recording start, 4) recording stop, 5) cleaning data and charts, 6) new window start, 7) browsing database.

The fields in section 2 represent signal charts: the waveform and spectra in different frequency scales. Fields e) to h) graphically visualize the instantaneous values of signal parameters during recording. Buttons in the bottom of the screen represent other operations related to sound recording. *Recording stop* (4) results in suspension of all actions performed by the program, i.e. recording, processing, and displaying samples. Pressing this button also causes averaging spectrum (LTAS), recalculating parameters and displaying their values in corresponding charts, as it is shown in Fig. 2. *Cleaning* (5) permanently deletes currently stored samples and results and enables new window to start, and (7) opens the database.

C. *Software and Hardware*

The software was written in C# in .NET framework using ASIO (Audio Stream Input/Output) audio interface. The relational database was built using SQLite software library that implements a self-contained, serverless, transactional SQL database engine. One row of the database represents a recorded sound file with the textual information, FFT coefficients and parameter values. Data are added to the database in a serial form, as an array of bytes. This is the core of the SQLite system which has a built-in binary large object storage class (BLOB). The serialization mechanism available in the environment .NET was used to transform data. The size of a single tuple in the database depends on the length of the sound file and number of related parameters. For example, a row containing a 59 second long audio recording, 320 sets of calculated parameters and 320 sets of FFT coefficients needs a capacity of 11 404 kilobytes (5151 kilobytes for audio). The contents of the database is presented in Table II.

Fig. 3 presents the implementation model of the MusicEar system. There are three operational layers: data, analysis and visualization. Transfer of information between different layers is carried out via the controller. Data can be downloaded from a database or via an ASIO driver (Audio Stream Input/Output - a computer sound card driver protocol) from a microphone and a sound cart.

A static class DBController was used as a data connector to communicate with the database module. It provides methods to

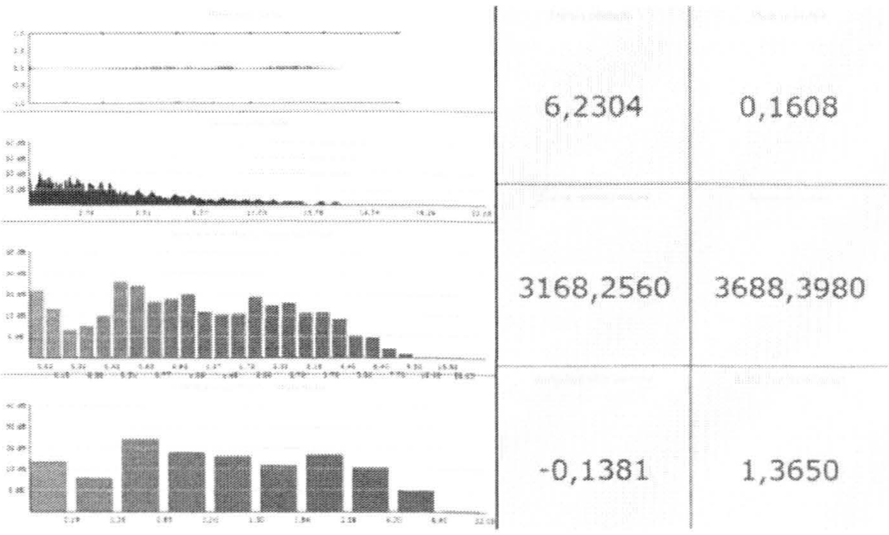


Figure 2. Long term averaged spectra and averaged values of parameters

enable "black box" data exchange - other parts of the system do not need to know how to get access to the database, only report the request to obtain data or to write data to the database.

The source code for SQLite is in the public domain [17]. There is no need to configure the database management system. Together with ADO.NET Entity Framework (ActiveX Data Objects for .NET is a set of computer software components that programmers can use to access data and data services) and a linking interface, the creation, modification, deletion, and access to the database is from the software environment.

Since MusicEar is a real-time system, it must fulfill the following minimal hardware requirements:

- Processor – dual-core, min 1,8 GHz,
- RAM – min. 2GB,
- Disc space – 20MB (for software only),
- Operating System – MS XP, Vista, Windows 7.

MusicEar operates with audio equipment: Audio interface E-MU Tracker Pre and the microphone DPA 4011.

TABLE II. DATABASE CONTENTS

Name	Data type	Comment
ID	numeric	Unique indentifier of the recording
Code	character string	Number of the violin, entered manually at the recording
Emblem	character string	Name of the violin, entered manually at the recording
Nationality	character string	Nationality (identifier) of the violinmaker, entered manually
Recording no	number	Recording number of the same violin
Date	date	Date of the recording generated automatically
Comments	Character string	Comment entered manually by the operator
Sound samples	byte array	Sound data,
FFT coefficienys	byte array	Spectrum samples
Parameters	byte array	values of all parameters

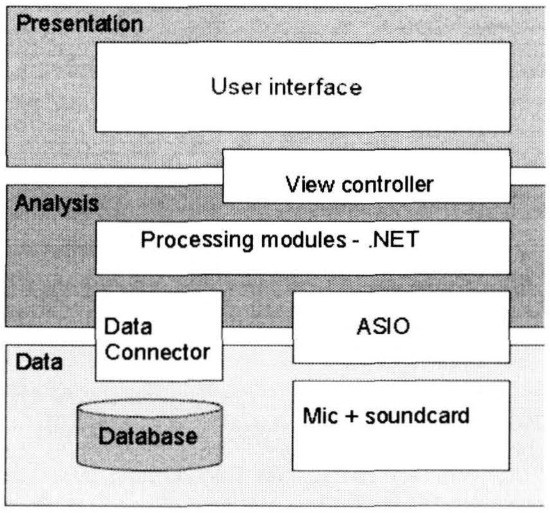


Figure 3. Implementation model of the MusicEar system

IV. MUSICEAR IN USE

First tests of MusicEar have been performed at the Academy of Music in Poznan, Faculty of Violinmaking. They enabled checking the system performance in the external environment with a violinist. These tests proved the concept of the system and its realization. The real-time condition was met and the system performance was satisfactory.

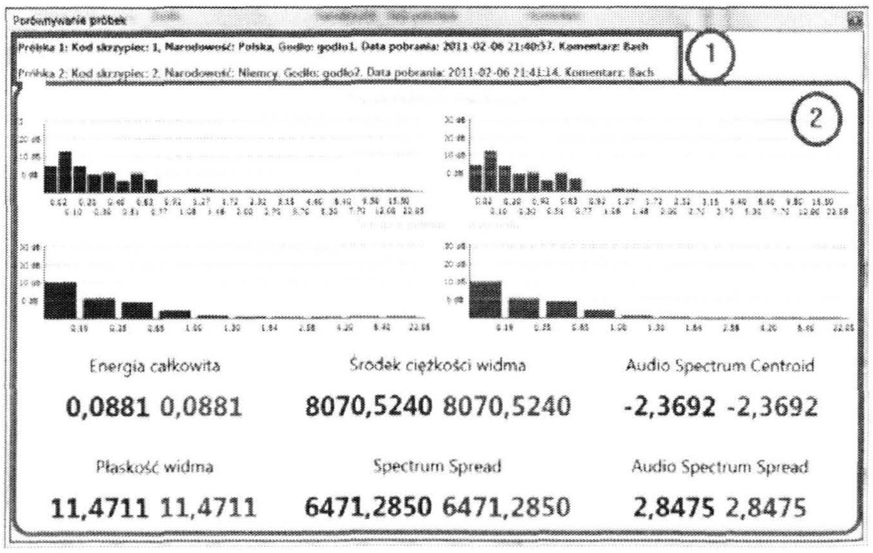
The subsequent tests were carried out during the International Henryk Wieniawski Violinmakers Competition held in Poznan in May 2011. The recordings took place during the competition auditions in the historic Groblicz chamber hall in the Museum of Musical Instruments. It was the first part of the final stage of the competition. Ten instruments competed at this stage. They were played five by five with a break in the middle of the audition. Each session started and finished with the reference violin, which, in case of this competition, was a Polish historical instrument of Bartlomiej Dankwart from 1602.

On the whole the MusicEar system worked well under time constraints of the competition. With a long recording there is a risk of RAM overload (note, that the recording and the calculated data are collected in RAM and transferred to the SQLite database as one tuple). For this reason one recording (violin No 5) was lost.

Observation of parameter values in real-time was very informative; the user could indicate the distinctive features of competing instruments while listening, and later in comparison window as presented in Fig.4. However multimodal analysis of violin sound by humans requires a large cognitive effort.

The sound material gathered during the competition brought a large amount of new data to analyze. The results of this analysis will not only characterize the competing instruments, but also will bring guidelines to develop and improve the MusicEar system.

As a preview of the results of the competing violins sound analysis two charts have been attached below. The figures feature information about nine competing violins and one historical reference violin. The numbers identifying violins are the same as during the competition. Fig. 5 presents the relationship between the energy and brightness and Fig. 6 presents the flatness.



Porównywanie próbek

Próbka 1: Kod skrzypiec: 1, Narodowość: Polska, Godło: godło1, Data pobrania: 2011-02-06 21:40:37, Komentarz: Bach

Próbka 2: Kod skrzypiec: 2, Narodowość: Niemcy, Godło: godło2, Data pobrania: 2011-02-06 21:41:14, Komentarz: Bach

Energia całkowita	Środek ciężkości widma	Audio Spectrum Centroid
0,0881 0,0881	8070,5240 8070,5240	-2,3692 -2,3692
Płaskość widma	**Spectrum Spread**	**Audio Spectrum Spread**
11,4711 11,4711	6471,2850 6471,2850	2,8475 2,8475

Figure 4. Sound comparison window

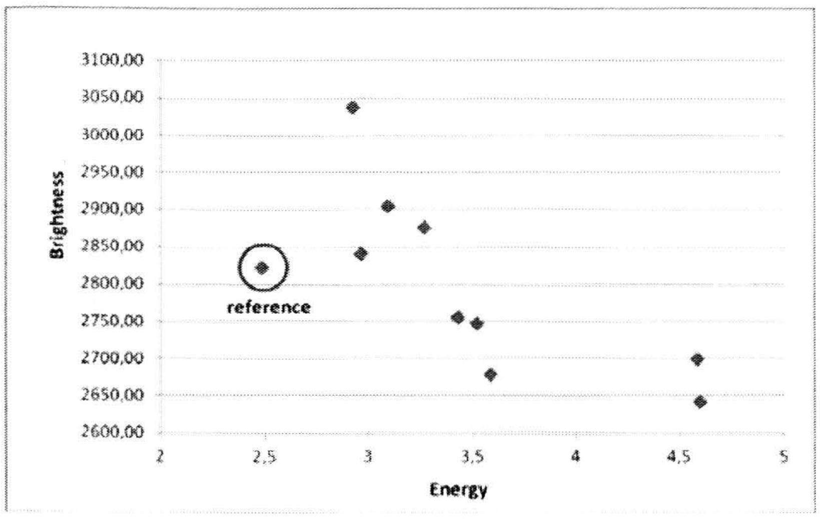


Figure 5. Relationship between the energy and brightness of competing violins

It is evident, that the historical violin differs from the rest of violins. It is less loud, less bright and has less tonal character than the contemporary instruments.

V. Conclusions

In the paper a report was made of the course of the project devoted to the implementation of a system for analysis, visualization and archiving of a violin sound in real time, which is designed to support the subjective assessment of the quality of the violin. The system, called MusicEar presents the waveform, instantaneous and long term averaged spectra in various frequency scales, and parameters related to the timbre and sound volume. The database collects the recorded sound, calculated spectra and parameters. Violins from the database may be compared already during the recording session. The first version of MusicEar was successfully tested in demanding conditions of the International Henryk Wieniawski Violinmakers Competition held in Poznan, Poland in 2011.

In this particular competition, the computer analysis results were not taken into account for jury decision and probably some time must elapse until this kind of analysis is approved for the violin quality assessment. Will it ever happen? Anyhow this type of software application may also be very useful for understanding the cognitive processes related to violin sound perception. It is hoped that next releases of the system will indicate similar instrument and will point the best.

The data gathered in the database of MusicEar are now further analyzed. In particular Duennnwald frequency regions are explored to find the reliable coefficients related to the quality of violin sound. Ultimately MusicEar is planned to become a real time decision support tool.

Acknowledgements

The authors would like to extend their thanks to Prof. Bartosz Bryla and Prof. Andrzej Lapa - Presidents of Henryk Wieniawski Musical Society in Poznań, to Andrzej Wituski – President of Henryk Wieniawski International Competitions and to Janusz Jaskulski – Curator of the Museum of Musical Instruments (Branch of the National Museum) in Poznan for making possible testing the MusicEar system and recording violins at the Competition and for their entire assistance.

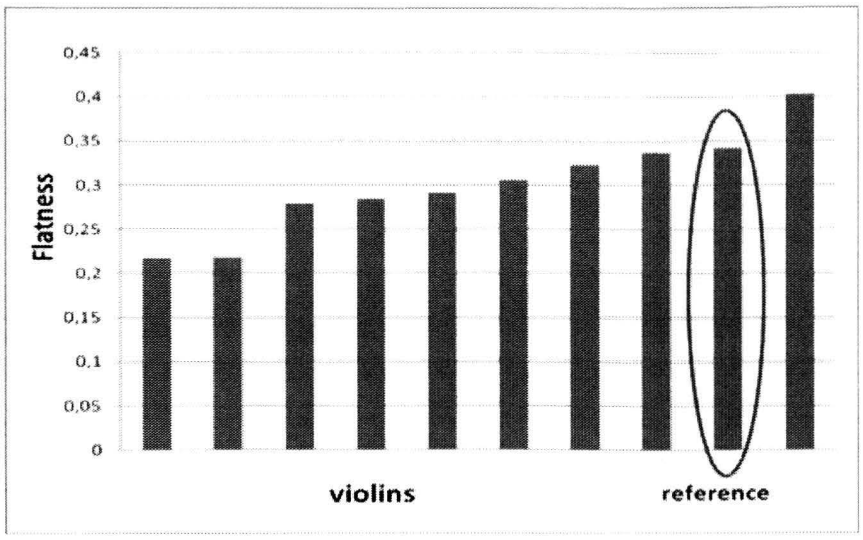


Figure 6. Flatness of competing violins

References

[1] A. Buen, "Comparing the sound of golden age and modern violins: Long-time-average spectra", VSA Papers, Vol. 1, No. 1, 2005, pp. 51-74.

[2] A. Buen, "Differences of sound spectra in violins by Stradivari and Guarneri del Gesu", Catgut Acoust. Soc. J., Vol. 4, No. 8 (Series 11), 2003, pp. 14-18,

[3] H. Duennwald, "Deduction of objective quality parameters on old and new violins" Catgut Acoust. Soc. J., Vol. 1 No 7, May 1991, pp. 1-5.

[4] A. Gabrielsson and E.V. Jansson, "Longtime-average spectra and rated qualities of twenty-two violins", Acustica; Vol. 42, pp. 47-55 (1979). Repr. in Res. Papers in Violin Acoustics 1975-1993, Vol. 1, C.M. Hutchins and V. Benade, Eds. (Acoust. Soc.Am., Melville, NY, 1997).

[5] Hutchins C.M. (Ed.), Benade V. (Ass.Ed.), Research Papers in Violin Acoustics 1975-1993, vol.1, vol.2, Acoust. Soc. of America Melville, NY, 1997.

[6] Hutchins C.M., A "History of Violin Research", JASA, 73, 1983, pp.1421-1432,

[7] ISO/IEC 15938-4, Information Technology, Multimedia, Content Description Interface, Part 4:Audio, 2001.

[8] E.V. Jansson, "Long-Time-Average-Spectra applied to analysis of music. Part III: A simple method for surveyable analysis of complex sound sources by means of a reverberation chamber", Acustica, vol. 34 no 5, 1976, pp. 275-280.

[9] B. Kostek, Perception-Based Data Processing in Acoustics, Springer-Verlag, GmbH, Berlin, Heidelberg, 2005.

[10] E. Łukasik, "AMATI-Multimedia Database of Violin Sounds", Proc Stockholm Music Acoustics Conference, KTH Stockholm, 2003, 79-82.

[11] S. S. Stevens, J. Volkmann, and E.B. Newman, "A scale for the measurement of the psychological magnitude pitch", Journal of the Acoustical Society of America 8 (3), pp. 185–190. 1937.

[12] A. Wieczorkowska, J. Zytkow, "Analysis of feature dependencies in sound description", Journal of Intelligent Information Systems, Vol. 20, No. 3, 2003, 285-302

[13] P. Wrzeciono, K. Marasek Violin Sound Quality: Expert Judgements and Objective Measurements" in: Z.W. Raś and A. Wieczorkowska (Eds.) "Advances in Music Information Retrieval", Studies in Computational Intelligence, Springer 2010, Vol. 274, pp. 237-260.

[14] E. Zwicker, Psychoakustik. Springer, Berlin, 1982

[15] http://www.svibs.com/products/ARTeMIS_Analyzer.aspx

[16] http://www.spectraplus.com/

[17] www.sqlite.org

Follow That Tune – Dynamic Time Warping Refinement for Query by Humming

Bartłomiej Stasiak
Institute of Information Technology, Lodz University of Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
Email: basta@ics.p.lodz.pl

Abstract — **Dynamic Time Warping is a standard algorithm used for matching time series irrespective of local tempo variations. This type of variability is inherent to audio input data obtained directly from users and, as such, it occurs in the context of Query-by-Humming interface to multimedia databases. Apart from the time-alignment problem, most of the known melodymatching approaches are also affected by a second issue of aligning the pitch between the query submitted by a user and the template. The query is usually in a different key and it may be simply sung out of tune, which needs some additional, sometimes computationally expensive processing and may not guarantee the success e.g. in the presence of pitch trend or accidental key changes.**

The method of *tune following*, proposed in this paper, enables to solve the pitch alignment problem in an adaptive way inspired by the human ability of ignoring typical errors occurring in sung melodies. The experimental validation performed on the database containing 4431 queries and over 5000 templates confirmed the enhancement introduced by the proposed algorithm in terms of the global recognition rate.

I. Introduction

The impressive diversity of methods and goals formulated in the area of Music Information Retrieval (MIR) reflects the intrinsic complexity of our perception of music and of the music itself. Out of the many research issues considered in the field, the problem of query specification for content-based music retrieval has been attracting significant attention for years. Among many proposed solutions, such as Query by Tapping, pitch contour specification with Parsons code or various forms of simplified musical notation, the Query by Singing/Humming (QbSH) interface is perhaps one of the most natural approaches to searching for a piece of music in multimedia databases.

The main issue in QbSH problem is basically *melody matching* where the melody is understood as a sequence of notes with given pitches and durations. Converting the user input into a sequence of pitch values, known as a *pitch vector*, is therefore a typical first step of processing. Many pitch detection algorithms (PDA) are available for this purpose [1][2][3], so a reliable representation may be usually obtained even in a relatively noisy environment. The potential problems involved here include the frequency resolution and precision of the PDA (usually of minor significance in the QbSH task), octave errors (may occasionally become an issue) and the imprecision of the sung query itself which is one of the main sources of confusion in practice.

The precise onset time and duration of a note are more difficult to be unambiguously determined. This is a point at which the approaches used for solving the QbSH problem may be roughly divided into two main groups.

1) Note-based Approaches: These methods aim at obtaining a reliable note segmentation with respect to the pitch and temporal parameters. Their biggest advantage is a compact representation allowing for efficient melody searching with string matching algorithms [4]. The methods proposed here include edit distance computation based on note insertion/deletion/replacement cost [5], transportation distances such as the Earth Mover Distance (EMD) [6] [7] and n-grams matching [8]. The note-based methods rely on the quality of the note segmentation stage which generally makes them potentially imprecise or dependent on the user adhering to a requirement of singing every note on a given syllable (e.g. "ta" or "da", cf. [5]).

2) Direct Matching: In these approaches the note segmentation problem is deliberately ignored and the pitch vectors are directly compared on a per-frame basis. High matching precision may be usually obtained in this way but at the cost of increased computational complexity [9][10]. Not only is the melody representation much longer than the sequence-of-notes form, but also variations of tempo in the sung query make the standard Euclidean distance between vectors inaccurate and a more sophisticated matching algorithm must be applied.

The method of choice for aligning the query with a template via a non-linear scaling of the time domain is known as Dynamic Time Warping (DTW). Proposed initially for isolated words recognition [12] it has been widely adopted in many other fields of artificial intelligence and signal processing.

One of the fundamental issues in a practical application of the DTW algorithm for melody matching is to obtain a key-invariant representation. The melody is defined by a sequence of relative pitches, so their absolute values are basically irrelevant. The user can sing a melody in any key, so all the notes may be shifted with respect to the template by the same interval which may result in a large value of the DTW distance, even for a perfectly sung query. In this paper a novel solution is proposed, in which the query is "tuned in" to the template via gradual decrease of the pitch difference between the two.

In the next section the principles of the DTW algorithm will be briefly presented along with a summary of previous works which influenced the development of the method in the context of QbSH and melody matching problems. Next, the proposed modification of the algorithm and the results of experiments demonstrating the obtained enhancement in recognition rate will be presented.

II. Basic Concepts

A. Previous Work

The problem of minimizing the distance between two time series which may vary in time or speed occurs naturally in numerous application areas. Early works of Itakura [11] and of Sakoe and Chiba [12] introduced the DTW as an effective solution in the speech processing task. The fundamental concepts laid out in [11][12] have been later used with slight modifications in many fields of artificial intelligence and data mining, including audio and video stream monitoring, bio-medical signal inspection, financial data analysis, human motion and gesture recognition [13][14]. The variants of the method include full sequence matching [12] and subsequence matching [13][15]. Efficient indexing techniques allowing to significantly reduce the searching time in large databases were introduced in [14] and applied in the Musical Information Retrieval context in [10]. The application of several variants of the DTW algorithm for the QbSH problem has been addressed in numerous works, including [9][15][16][17][18].

B. The Fundamentals of Dynamic Time Warping

Let q_j denote the pitch value in the j-th frame of the query pitch vector $\boldsymbol{q}$, where $j = 1, 2, ..., J$. Similarly, t_i represents the i-th frame of the template $\boldsymbol{t}$, where $i = 1, 2, ..., I$. The Euclidean distance between the two:

$$d_{\mathrm{Euclid}}(\boldsymbol{q}, \boldsymbol{t}) = \sqrt{\sum_i |q_i - t_i|^2} \, , \qquad (1)$$

may be computed only if the size of the two vectors is the same, which is typically not the case. Moreover, reinterpolating the sequences linearly to the same length may not be sufficient in the presence of local tempo variations (Fig. 1).

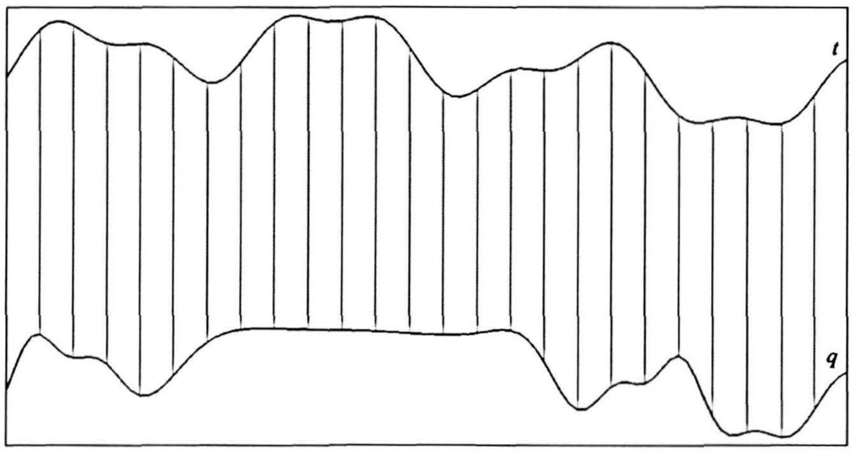


Fig. 1. Sequence matching with the Euclidean distance

The solution is to scale the time domain of the sequences with a proper *warping function* so that the corresponding frames are properly matched (Fig. 2). The warping function may be represented on the i-j plane by a path, i.e. a sequence

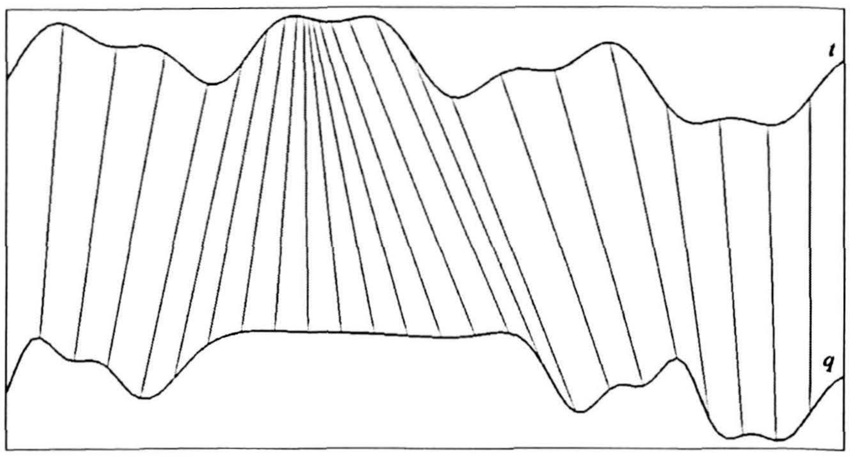


Fig. 2. Sequence matching after non-linear rescaling

of points $c(1), c(2), ..., c(K)$, where $c(k) = (i(k), j(k))$ (Fig. 3). Every path is assigned a cost:

$$E = \sum_{k=1}^{K} d(c(k)) \, , \qquad (2)$$

where the cost of matching an individual point $c(k)$ may be defined as:

$$d(c(k)) = d(i, j) = |q_{j(k)} - t_{i(k)}| \, . \qquad (3)$$

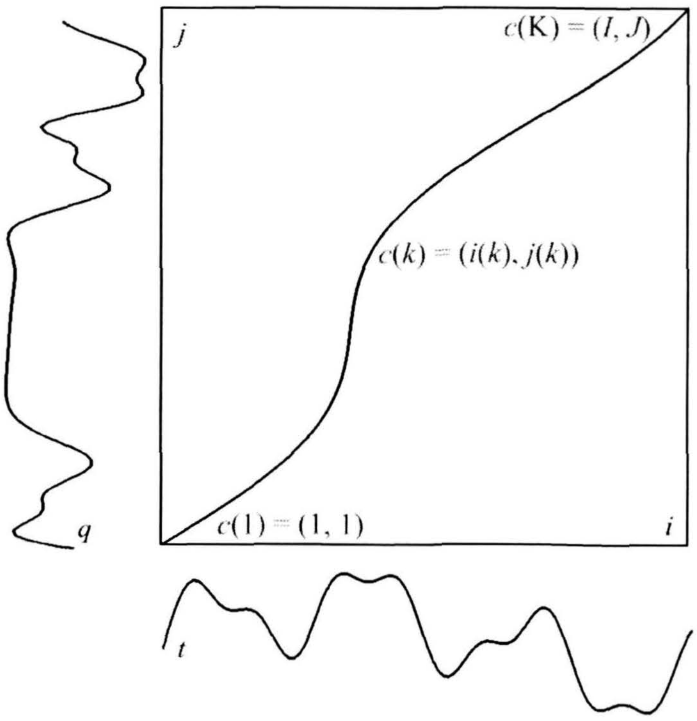


Fig. 3. The warping function

The DTW algorithm, finding the optimal path in the sense of minimization of eq. (2), is based on the dynamic programming (DP) principle. The i-j plane is represented as a two-dimensional array g. Every element $g[i, j]$ is assigned a minimal cost of reaching the point (i, j) from the beginning point $c(1) = (1, 1)$:

$$\underset{\substack{i=1,2,...,I \\ j=1,2,...,J}}{\forall} \quad g[i,j] = d(i,j) + \min \begin{cases} g[i, j-1] \\ g[i-1, j-1] \\ g[i-1, j] \end{cases} \qquad (4)$$

with the boundary conditions:

$$\begin{aligned} g[0,0] &= 0 \,, \\ g[0,j] &= \infty, \text{ for } j=1,2,...,J \,, \\ g[i,0] &= \infty, \text{ for } i=1,2,...,I \,. \end{aligned} \qquad (5)$$

After computing all the values of the array g, the total cost of the optimal path is found in $g[I,J]$. This value is typically multiplied by $(I+J)^{-1}$ to allow comparisons between queries of different lengths.

The DP-equation (4) is a simple variant most often found in literature [13][14]. Several more sophisticated variants incorporating local slope constraints and weighting coefficients were initially proposed by Sakoe and Chiba [12]. Global constraints in the form of the *Sakoe and Chiba band* [12] or *Itakura parallelogram* [11] are also often applied (Fig. 4). The general role of the constraints is to limit the area of the i-j plane under consideration in order to speed up computations and to reduce the risk of "pathological warping" of the sequences [14]. Global constraints play also fundamental role in efficient indexing techniques introduced by Keogh in [14].

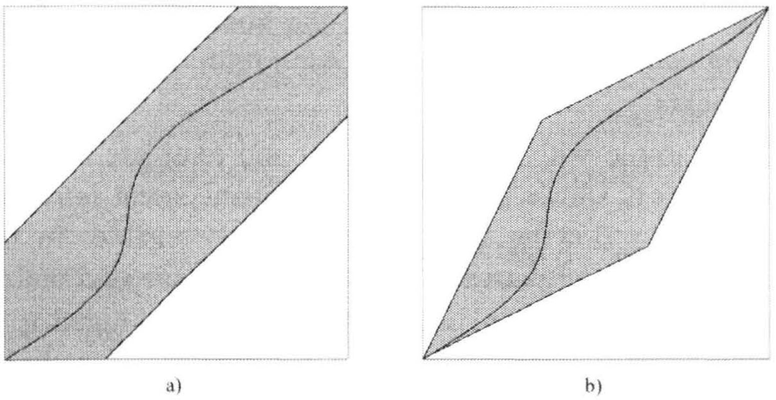


Fig. 4. DTW global constraints: a) Sakoe and Chiba band, b) Itakura parallelogram.

The boundary conditions may be modified to allow for a situation when only a fragment of one sequence is to be matched against the second one. This is generally a subsequence matching problem in which the compared sequences may not start at the same position and/or end at the same position [13].

C. Melody Matching

In a typical approach, the query sung by a user is matched against a database consisting of a collection of MIDI files. The templates from the database are converted, similarly to the query, to the form of pitch vectors, expressed in the MIDI note numbers rather than as frequency values in Hz. The conversion is straightforward in the case of the MIDI files and always yields unambiguous results. On the other hand, the query pitch vectors often need some clean-up to decrease the influence of noise, octave errors, etc. and they generally represent the intended melody only approximately. Many users sing out of tune and they cannot sing with sufficient precision, especially in case of bigger intervals [19].

The general problem which is addressed in this work is how to match a melody sung in a different key than in the template. There exist several approaches to deal with this issue. Many researchers use a simple method of subtracting the mean pitch from the whole sequence [18]. The problem occurs when the melody represented in a query is only a part of the template, or vice versa, in which case subtracting the mean is of no use.

In a different approach the melody may be represented in the form of relative changes of consecutive pitches (differential/delta representation) [20]. This eliminates the problem but representing raw pitch vectors in this form often yields poor results. In this case the MIDI-based templates consist mostly of zeroes with non-zero values only at the points of note transitions. On the other hand, the note transitions in a query may be spread over several frames which makes the true comparison impossible.

An effective alternative may be to repeat the matching procedure several times with different transpositions of the query pitch vector. The query may be transposed by e.g. all possible number of semitones within the octave [16] or from -5 to $+5$ semitones in half-of-the-semitone steps [20]. Various numbers of repetitions may be considered but in any way this is clearly a brute-force approach which increases the computational complexity significantly. Another problem, which is still not solved, is that the transposition may appear within the query when the user fails to sing an interval (usually a greater one) precisely and continues in a different key.

A solution proposed in this work is to try to follow the melody of the template by gradually decreasing the difference between the query and the template. This is intended to resemble the way in which humans follow the known melody irrespective of pitch inaccuracies and key changes.

III. The Proposed Algorithm

The input query pitch vector $\boldsymbol{q}_{\mathrm{raw}}$ is obtained from audio data sampled at 8kHz, with the non-overlapped frame size of 256 samples. It is first preprocessed in order to obtain a smooth melody line without large jumps and unvoiced fragments. The preprocessing includes the following steps:

1) The leading and trailing unvoiced fragments, denoted by the pitch detection algorithm as "0", are removed.
2) The median of the remaining data is computed and all the pitch values distant from the median by more than a given threshold T_1 are marked as unvoiced i.e. set to zero. This may help in case of poor quality of the input data resulting from noise or from errors introduced by the pitch detection algorithm. The quality of the database used in the experiments made this correction necessary in 1% of the queries for $T_1 = 24$ semitones.
3) For the same reason the maximum jump between two consecutive frames can not exceed a threshold T_2. Setting $T_2 = 14$ semitones resulted in 3.8% corrected files.
4) Every unvoiced frame is set to the pitch value of the last voiced frame. In this way one continuous melody is obtained, without any breaks resulting from breathing or

articulation. It should be noted that this operation also leads to rejecting some potentially useful information about the rhythm and beat.

5) Median filter with the size of 9 frames is applied to smooth the pitch contour. Preliminary experiments showed that it enhances the recognition results significantly.

The smoothed query pitch vector $\boldsymbol{q}$ is then compared with all the templates from the database. For every template $\boldsymbol{t}$ the pitch difference d_{beg} between the beginnings of the query and the template is computed and then subtracted from all the elements of the query pitch vector:

$$\underset{j=1,2,\ldots,J}{\forall} q_j := q_j - d_{\text{beg}} \,, \tag{6}$$

where J is the length of $\boldsymbol{q}$ after preprocessing.

This makes both sequences start in the same key. In practice, the value of d_{beg} is computed as:

$$d_{\text{beg}} = \frac{q_2 + q_3}{2} - \frac{t_2 + t_3}{2} \,. \tag{7}$$

The first pitch value may be unreliable, so it is rejected and the mean of the next two is taken into account.

As the database used for the experiments contained only queries sung from the beginning, this procedure enabled to obtain good matching results with the standard DTW algorithm described in section II-B. On the other hand, the queries from the database often ended in arbitrary positions with respect to the template sequences, so using the arithmetic mean computed for *all* the values of $\boldsymbol{q}$ and $\boldsymbol{t}$ instead of d_{beg} in (6) yielded poor results.

In a separate set of preliminary experiments the influence of DTW constraints on the recognition results has been tested. It has been found that setting the slope constraint condition $P = 1/2$, as defined in [12], yielded the best results.

Having the beginning of the query shifted properly along the frequency axis, one have to deal with transpositions possibly occurring later in the course of the query (Fig. 5). For this purpose the standard DTW procedure is applied first to find the warping function aligning the query and the current template. Going along the path on the i-j plane defined by the warping function, the procedure defined by the block diagram in Fig. 6 is applied. The resulting signal $\hat{\boldsymbol{q}}$ is a version of $\boldsymbol{q}$ modified to follow the pitch values defined by the template $\boldsymbol{t}$. This process is controlled by the parameter $\alpha \in (0, 1]$. The greater the value of α, the faster will the pitch of the query be aligned with the template. The final value of $\alpha = 0.05$ was used in the experiments.

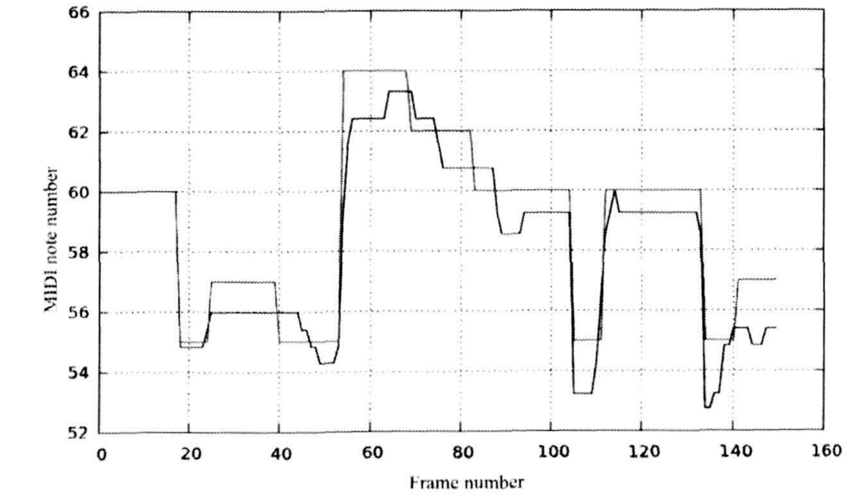


Fig. 5. Example of a transposition (*Old McDonald had a farm*). The first 17 samples of the template (light) and the median-filtered query (dark) are in tune. Most of the remaining part of the query is one – two semitones below the template.

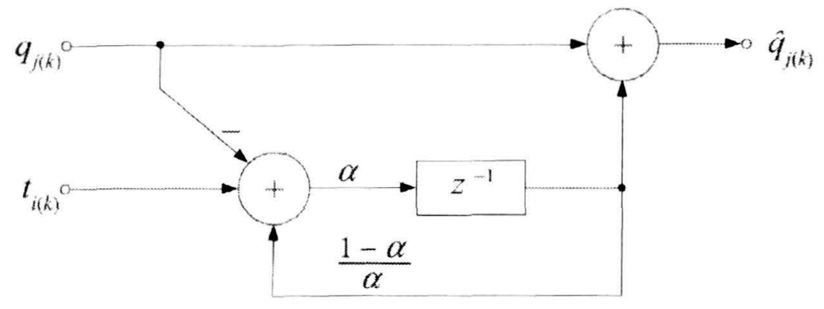


Fig. 6. The block diagram of the tune follower.

The example with the same query and template sequences as in Fig. 5 is shown in Fig. 7. The enhancement introduced by the tune-following procedure is clearly visible. In most places the distance between the sequences decreased and two fragments in the second half of the query got tuned to the template exactly. It should be noted that both Fig. 5, and Fig. 7 present the aligned, i.e. time-warped version of the sequences.

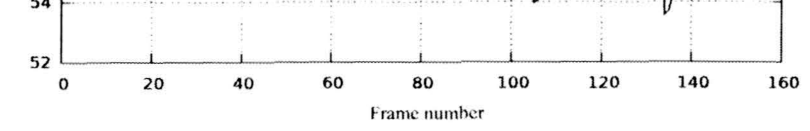


Fig. 7. The result of application of the tune follower ($\alpha = 0.1$).

The final matching cost is then computed for the sequence $\hat{\boldsymbol{q}}$ with formula (2). One important thing that should be noted here is that although this cost is lower in comparison to the standard DTW algorithm for the matching template, it can also be lower for the non-matching ones. The fundamental question is whether the proposed tune-following procedure is able to make the matching template win easier in the competition with the others despite the fact that all of them may benefit from its application. In the following section the test results supporting positive answer to this question have been presented.

IV. Experimental Results

A. Database

The publicly available datasets, used in the MIREX 2011 *Query by Singing/Humming* evaluation task [21], have been chosen to verify the proposed solution. Roger Jang's MIR-QBSH corpus [22] consists of a collection of 48 popular songs (ground-truth MIDI files) to be matched against 4431 queries sung by about 200 subjects. The 48 ground-truth files are mixed with 5274 "noise" files from Essen collection [23].

B. Testing Procedure and Results

Each of the 4431 queries was compared with all of the 48+5274 template files, one of which was the correct one. For every query q, all the templates were ordered by their DTW distance from q . According to the rules used in the MIREX evaluation, the search was treated as successful when the correct template was among the top 10 results. The obtained results are presented in Table I.

TABLE I
Total Number of Recognized Queries

	Top Ten Score	Best Hit Score	δ
DTW	3077 (69.44%)	2109 (47.60%)	0.39532
DTW + Tune Follower	3332 (75.20%)	2455 (55.41%)	0.42975

Additionally, the number of cases when the correct template was the first one on the list of DTW distances was also recorded (the *Best Hit Score* column). The last column displays the mean relative difference between the first and the second file on the list:

$$\delta = \frac{1}{N} \sum_{n=1}^{N} \frac{E_2^{(n)} - E_1^{(n)}}{E_1^{(n)}} , \qquad (8)$$

where the sum is computed only over those N queries for which the best hit was the correct one ($N = 2109$ or 2455, respectively). The value $E_p^{(n)}$ denotes the DTW matching cost for the n-th query and the template located at p-th position on the list, i.e. the value $E_1^{(n)}$ represents the score of the template best matching the n-th query (naturally, $E_1^{(n)} < E_2^{(n)}$).

The character of changes introduced by the proposed algorithm may be better assessed on an example of a single query shown in Table II. The correct template was found to be the closest to the query, both with and without the tune follower (all of the remaining files come from the Essen collection). It may be however observed that the DTW distance of the first template decreased significantly, from 544.56 to 382.80, while the second template remained almost equally distant from the query (659.72 vs. 675.03). Application of the tune follower reordered the list and introduced some changes in the top-ten matching templates (e.g. file Q0095.pv appeared and Q0082.pv was removed).

TABLE II
Results for a Single Query
Year: *2003*, Person: *00011*, File: *00020.pv* (*Happy Birthday*)

	DTW		DTW + Tune Follower	
No	Template	DTW Distance	Template	DTW Distance
1	00020.pv	544.56	00020.pv	382.80
2	V0003E.pv	675.03	V0003E.pv	659.72
3	E0820.pv	731.27	E0820.pv	672.47
4	A0302.pv	752.51	Q0075P.pv	678.65
5	Q0114N.pv	814.22	Q0095.pv	697.24
6	Q0082.pv	825.53	A0302.pv	712.63
7	Q1102J.pv	830.32	Q0114K.pv	712.67
8	Q0080B.pv	840.34	Q0137E.pv	734.59
9	Q0080A.pv	849.02	Q2079J.pv	738.94
10	E0110B.pv	876.25	Q0048C.pv	745.74

C. Discussion

The presented results consistently show that the proposed tune-following procedure may have a positive influence on the DTW-based melody search. Although it is true that it generally makes the matching cost smaller for most of the templates, one can expect that this decrease will be more significant in the case of the correct template than for all the non-matching ones (Table II).

This may result from the effect of accumulation of the corrections for consecutive notes. For example, when the pitch of a note sung by a user is too low with respect to the correct template then it is gradually increased by our procedure until it reaches the right tune, provided that the note is long enough. If it is relatively short, it is at least partially corrected. In either case, if the note was sung too low, then it is probable that the pitch of the next note will also be too low in which case it will get corrected immediately or – at least – faster. This effect may be observed e.g. when comparing Fig. 5 and Fig. 7. The pitch discrepancy in frames 105–110 is made significantly smaller due to correction which occurred in the previous frames.

This type of correspondence between the signs of the pitch differences in consecutive notes cannot be generally expected when comparing a query with a non-matching template. Correcting one note may result in increasing the initial difference between the next note and the template. This may even result in increasing the total matching cost, although for long notes and infrequent pitch changes the tune follower will make the query closer to most of the templates.

Further investigation revealed that the exact number of cases when the standard DTW failed to put the correct template on the first place and at the same time the proposed solution managed to do so, was equal to 493. Yet in 147 cases the opposite was true, i.e. the correct template disappeared from the first position when the tune follower was turned on. These figures are definitely dependent on the parameter α of the tune follower. Finding the optimal value for α needs some additional tests and close inspection of those 147 cases.

In general, the proposed solution enables to efficiently refine the results without computationally complex methods such as repeating the DTW for all possible transpositions [16]. It should be noted that it can be used independently on efficient indexing techniques [10][14] or note-based approximate algorithms [17] to increase the speed and reliability of a QBSH-based search engine.

V. Conclusion and future works

In this work a modification of the Dynamic Time Warping procedure have been proposed to enhance the results of melody matching in the Query by Humming problem. The modification is inspired by the human ability to match melodies irrespective of the key and pitch inaccuracies. It may be stated that the proposed tune-following procedure plays a similar role for pitch alignment as the DTW does for the case of time alignment and thus it may be seen as a frequency-domain complement to DTW. Similarly, while the DTW decreases the matching cost with respect to the Euclidean distance, the tune-following procedure decreases it even more, with respect to the DTW alone. Although the distance is lower both for the matching and non-matching templates, the presented experimental results clearly demonstrated the superiority of the proposed solution in terms of recognition rate and separation between the matching and non-matching templates.

The concept of tune-following will be further investigated in future works. Apart from the issue of parameter settings and possible modifications of the presented procedure itself, it should be noted that it is currently being applied to the already time-aligned sequences, i.e. after the DTW algorithm. It is however possible to integrate the two and modify the pitch adaptively during the dynamic programming optimization of the path cost. This would enable to obtain a different warping function in some cases and, possibly, to match more imprecisely sung queries. However, it seems unclear if this would lead to the overall recognition rate improvement – some further research is hence necessary here.

The generalization of the proposed method to subsequence matching problem would also be of great practical importance. It would eventually enable to construct a flexible hybrid system incorporating several methods, both direct and note-based, that would benefit from the tune-following algorithm to offer enhanced results in shorter time.

References

[1] M. Dziubiński, and B. Kostek, *High accuracy and octave error immune pitch detection algorithms*. Archives of Acoustics, Vol. 29, No. 1, 1-21, 2004

[2] D. Gerhard, *Pitch Extraction and Fundamental Frequency: History and Current Techniques*. Technical Report TR-CS 2003-06, Dept. of Computer Science, University of Regina, 2003

[3] P. Boersma, *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. IFA Proceedings 17, 1993

[4] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, *Query By Humming Musical Information Retrieval in an Audio Database*. ACM Multimedia, pp. 231–236, 1995

[5] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson and S. J. Cunningham, *Towards the digital music library: tune retrieval from acoustic input*. Proc. ACM Digital Libraries, pp. 11–18, 1996

[6] R. Typke, F. Wiering, and R. C. Veltkamp, *Transportation distances and human perception of melodic similarity*. Musicae Scientiae, Discussion Forum 4A, pp. 153–181, 2007

[7] S. Huang, L. Wang, S. Hu, H. Jiang, and B. Xu, *Query by humming via multiscale transportation distance in random query occurrence context*. IEEE Int. Conf. on Multimedia and Expo, pp. 1225–1228, 2008

[8] A. Uitdenbogerd and J. Zobel, *Melodic matching techniques for large databases*. In ACM Multimedia 99, pp. 57–66, 1999

[9] J.-S. Jang, and H.-R. Lee, *Hierarchical Filtering Method for Content-based Music Retrieval via Acoustic Input*. Proc. ACM Multimedia, pp. 401–410, 2001

[10] Y. Zhu, and D. Shasha, *Warping indexes with envelope transforms for query by humming*. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 181–192, 2003

[11] F. Itakura, *Minimum prediction residual principle applied to speech recognition*. IEEE Trans. on Acoustics, Speech and Signal Processing, pp. 67–72, 1975

[12] H. Sakoe, and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition*. IEEE Trans. on Acoustics, Speech and Signal Processing, pp. 43–49, 1978

[13] Y. Sakurai, Ch. Faloutsos, and M. Yamamuro, *Stream Monitoring under the Time Warping Distance*. Research showcase, Carnegie Mellon University, http://repository.cmu.edu/compsci/529, 2007

[14] E. Keogh, *Exact indexing of dynamic time warping*. In 28th International Conference on Very Large Data Bases, pp. 406–417, 2002

[15] J. Lijffijt, P. Papapetrou, J. Hollmen, and V. Athitsos, *Benchmarking dynamic time warping for music retrieval*. Proc. of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments (PETRA '10), No. 59, 2010

[16] H.-M Yu, W.-H. Tsai, and H.-M. Wang, *A Query-by-Singing System for Retrieving Karaoke Music*. IEEE Transactions on Multimedia, Vol. 10 (8), pp. 1626–1637, 2008

[17] L. Wang, S. Huang, S. Hu, J. Laing, and B. Xu, *An effective and efficient method for query by humming system based on multi-similarity measurement fusion*. Int. Conf. on Audio, Language and Image Processing, pp. 471–475, 2008

[18] W. Jeon, and Ch. Ma, *Efficient search of music pitch contours using wavelet transforms and segmented dynamic time warping*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2304 - 2307, 2011

[19] J. Yang, J. Liu, and W.-Q. Zhang, *A Fast Query by Humming System Based on Notes*. Interspeech 2010, pp. 2898–2901, 2010

[20] http://www.music-ir.org/mirex/abstracts/2011/JSSLP1.pdf

[21] http://www.music-ir.org/mirex/wiki/2011:MIREX_Home

[22] mirlab.org/dataSet/public/MIR-QBSH-corpus.rar

[23] www.esac-data.org

Melody Recognition System

Katarzyna Adamska
Institute of Electronics
Technical University of Łódź
Łódź, Poland
kate_adams@windowslive.com

Paweł Pełczyński
Institute of Electronics
Technical University of Łódź
Łódź, Poland
pawel.pelczynski@p.lodz.pl

Abstract **— The purpose of the presented research was the development of a melody search system that would allow to find a song in a music database, based on humming the tune of a known song fragment. The melody recognition in the developed system is based on comparing vectors of voice pitch values. The best match of the humming pitch vector against all the recordings in the database is searched. An original frequency and time scaling approach was implemented to improve recognition accuracy. The system performance was tested with the use of live humming recordings of four volunteers with a small database.**

Keywords — melody search, note frequency, query by humming, melodic contour.

I. Introduction

The task of recognizing the melody is one of the issues involved in the processing and storage of data related to music. The process of computerization of music helps not only people professionally involved in music, such as composers, instrumentalists or vocalists, but also provides new opportunities for the recipient of a musical work - the listener.

Currently, the most popular method of search for songs in musical databases is the so called "search by text" [1]. The searched text may be the name of the song, composer, artist or any fragment of the lyrics. It turns out, that it is not a satisfactory solution, when one remembers only a fragment of a song's melody. A melody recognition system allows the user to simply hum a song fragment, then treat it as a query to the database. As a result a list of tracks, that most fit to the hummed fragment, is returned. This is a special case of query by example, called the *Query by Humming* (QbH) [2]. For different recordings of the same musical work the only relatively constant feature of the song is the melody.

The melody is a fundamental part of any musical work. A song may consist of several tunes, and usually one of them is the main melody while the rest form the musical accompaniment. A melody consists of several notes occurring in a particular order. For each note one can specify the frequency and duration. Note representation of a song can be easily implemented in a musical database by direct note writing or can be retrieved from MIDI files. A properly hummed fragment of a song allows to recognize its melody for a human listener. Many QbH systems rely on note recognition. In *Note Interval Matching* [3] melodies are treated as strings. They are aligned to obtain the best similarity of two melodies with the use of dynamic programming. Frequency variation between consecutive notes serves as the measure of similarity. Another, well known approach, *N-Gram Matching* [4], is also based on relative note frequency, but it is quantized and an exact match is searched. A combination of the above two techniques is often used in a *Two-stage Search*: N-Gram Matching allows for rejection of the worst guesses and Note Interval Matching refines the result in a smaller database. Unfortunately, automatic note recognition is not an easy task [5]. Lack of clear boundaries between notes in a real humming signal, both in amplitude and frequency, makes automatic note recognition prone to errors due to inaccurate singing and background noise. To avoid the above problems the authors have decided to develop and investigate a melody search system without note recognition, as in *Melodic Contour Matching* techniques [4]. Instead of note symbols a vector of fundamental frequencies estimated in short time intervals, called "melodic contour", is produced. This vector is matched to similar data structures, stored in the database. The best match should occur for the proper piece of music, allowing correct recognition. Typically, *Dynamic Time Warping* is used to find the best match of two melodic contours. This algorithm is time consuming, which motivated the authors to replace it with fine scaling of tempo and pitch of the hummed melody.

II. The concept of the melody regognition system

The presented project consists of three parts. The first is the analysis of the humming recording to identify the fundamental frequency for small time intervals and to find its variability in time. The second part relates to the extraction of the melodic contour of the songs in a database. The third part is the matching algorithm, which allows to search hummed tone sequence in the database of songs. The block diagram of the developed system is shown in Fig. 1. Signal processing and analysis procedures are represented with ovals, database records are enclosed in cylinders, and the produced data vectors are in rectangles. The research work is focused on the development and testing of these algorithms, it does not cover optimization problems of the music database search. To simplify the tests a MIDI representation of songs was chosen. It required the extraction of notes and its conversion to melodic contour of each record.

III. Extraction a melodic contour of humming sound signal

Melody matching in the proposed approach is based on the only one signal feature - the frequency of the first signal

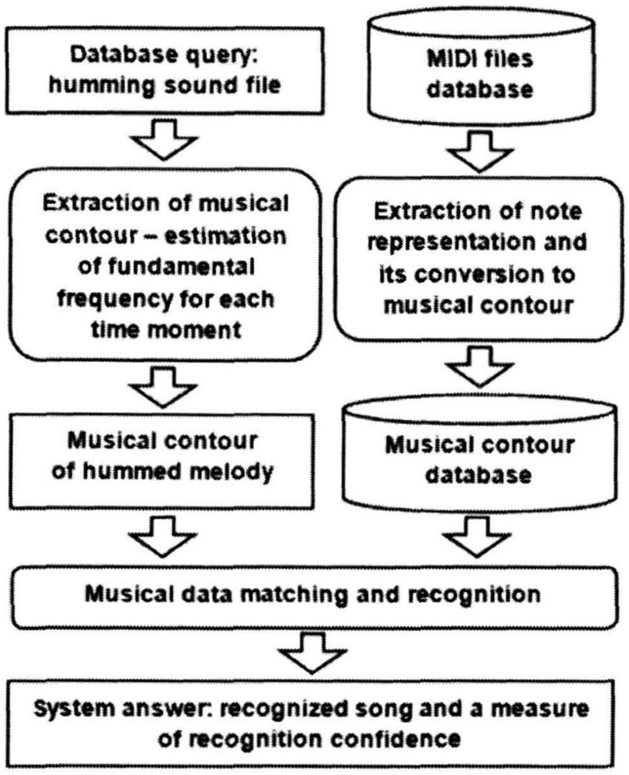

Database query: humming sound file
MIDI files database
Extraction of musical contour – estimation of fundamental frequency for each time moment
Extraction of note representation and its conversion to musical contour
Musical contour of hummed melody
Musical contour database
Musical data matching and recognition
System answer: recognized song and a measure of recognition confidence

Figure 1. A Schematic diagram of the developed melody recognition system

harmonic, known also as the pitch. In general pitch is the subjectively perceived frequency, not always representing a real one [7]. Fortunately a humming is a periodic signal with the first harmonic of considerable amplitude. Thus, pitch estimation algorithms can be applied in the stage of feature extraction. A lot of methods for pitch estimation exist. The most known are based on an Autocorrelation Function (AF), Cepstrum, Average Magnitude Differential Function (AMDF) or Comb Transformation, [8],[9].

The authors chose the algorithm based on the accurate autocorrelation method [10]. This method is more accurate and robust, than methods based on cepstrum or filtering. It was implemented in Praat sound processing and analysis environment [11], that can be invoked as a command line application from other programs. The main parameters of the utilized algorithm (function: Sound To Pitch) were set as follows:

- Time step – 10ms,
- Pitch floor – 75Hz,
- Max number of candidates – 15.

Other parameters were set to their default values. Sound analysis performed by "Sound To Pitch" function produces a vector of frequency samples, that approximates the melodic contour of a given musical work (Fig. 2.). The obtained vector was then compared with musical database. The developed matching algorithm is described in section V.

IV. Extraction of the melodic contour from MIDI files

Musical Instrument Digital Interface (MIDI) is a standard which allows collaboration between musical instruments, or between a computer and musical instrument. Communication between devices is realized using the MIDI protocol. This protocol is a fixed set of messages sent between devices. These messages equipped with time stamp form commands, that are stored in files, called MIDI files. MIDI commands can control an electronic, musical instrument or a synthesizer, that is often integrated in a sound card. A single MIDI file can be used to control up to 16 synthesizers – they are multipath. Each path carries the information about the played note numbers and their durations, which gives a synthetic representation of a melodic contour.

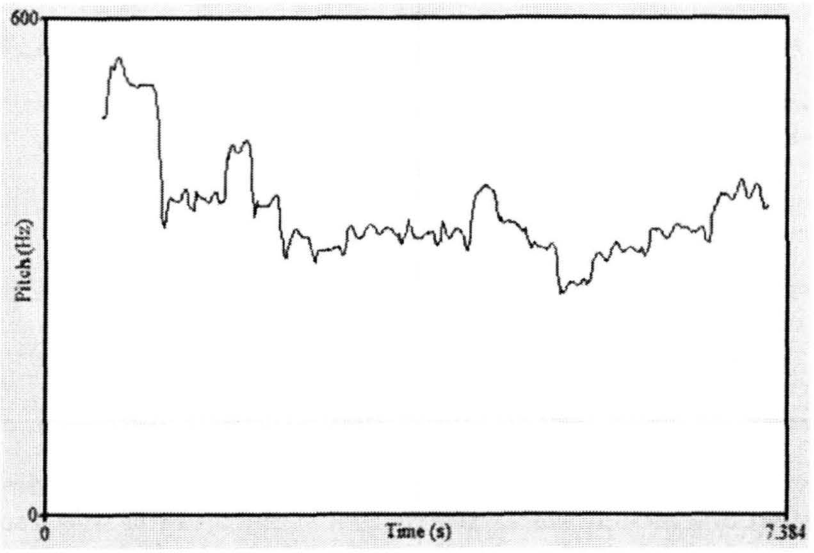


Figure 2. An example of the pitch vector extracted from a sample humming recording

A specialized converter from a MIDI file to sampled musical contour was written in Java. It is capable of extracting a sequence of desired commands and converting them into a series of frequency values. The "javax.sound.midi" packet was used to read and perform an analysis of MIDI files. A conversion from note MIDI numbers to frequency required a knowledge of the frequency mapping. It is described by the following formula:

$$f_{note}=2^{x/12}\cdot f_{ref} \tag{1}$$

where:

$$x=n_{MIDI}-n_{ref\,MIDI} \tag{2}$$

Reference note is A4, its frequency f_{ref}=440Hz, and MIDI number $n_{ref\,MIDI}$=69. The coefficient $2^{1/12}$ is a halftone frequency interval.

Producing samples of melodic contour required both a decision of sampling period and measuring a time in MIDI. A choice of period was straightforward: it was the same as the time step in the Praat "Sound To Pitch" procedure. An estimation of time in MIDI depends on two parameters stored in the file: the tempo expressed in "beats per minute", and the resolution expressed in "tics per beat". Each event, such as the beginning and the end of a note, is expressed as a tick number, which is converted to discrete time based on the sampling period. Building of the melodic contour vector starts from determining its length and filling all the elements with zeros. Then, for each note, a proper range of vector elements is filled with note frequency. An example of extracted melodic contour can be seen in Fig. 3.a).

V. Melodic contour matching and song recognition

A recognition of a melody cannot be simply a result of comparing the unknown melodic contour with the contours stored in the database due to their different lengths and lack of time synchronization. It should be obvious that there is no need to hum the whole song and that the hummed fragment does not have to start from the beginning of the song. In these conditions a recognition is a result of finding the best match of the unknown melodic contour to the fragment of one of the contours from the musical database. Assuming, that the length of the unknown contour vector is not higher than the lengths of all the vectors in the database, an error of vector match to pattern *p* for a given time shift *i* can be defined as a Root-Mean-Square Error (*RMSE*):

$$e_p(i)=\{ [\Sigma_{j=0}^{N-1}(x(j)-p(j+i)^2] / N \}^{1/2}, \quad 0 \le i \le M-N \tag{3}$$

where:
x – unknown melodic contour vector of length N,
p – p-th melodic contour vector from database of length M.

Both vectors and a plot of error e_p are shown in Fig. 3.

Using *RMSE* as the measure of the matching accuracy makes the result independent of vector length and allows to express it in frequency units – [Hz]. The best match error me_p to the pattern p over all time shifts is the minimum of $e_p(i)$:

$$me_p = min\{ e_p(i) \}, \quad 0 \le i \le M-N \tag{4}$$

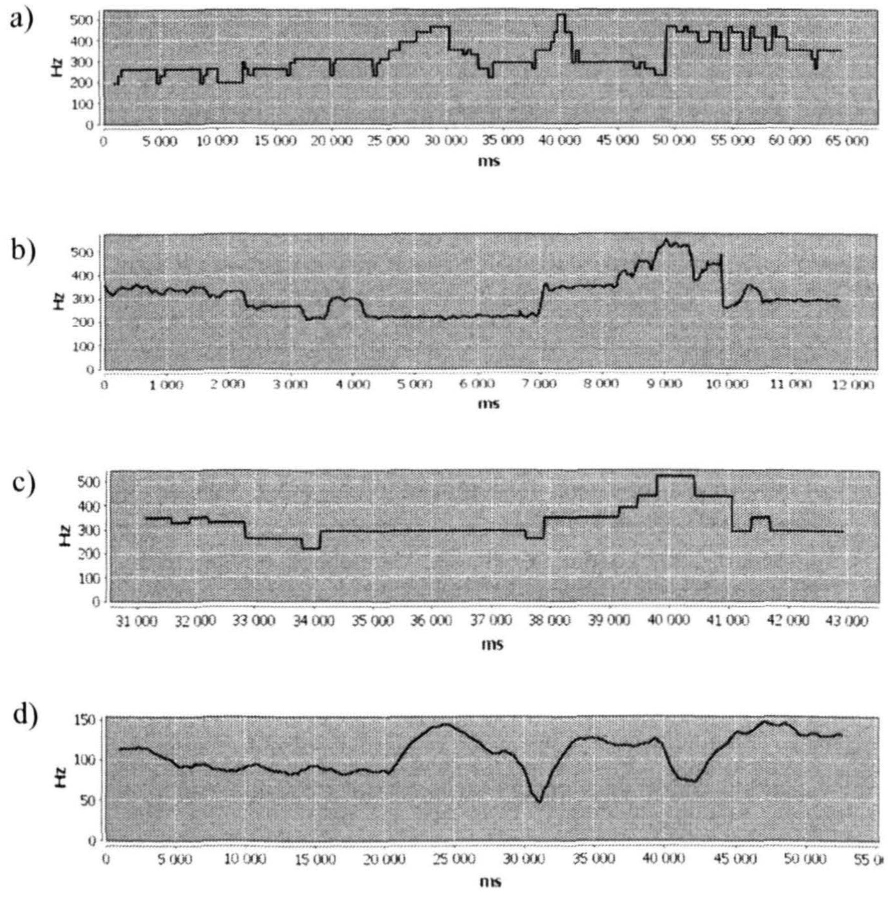


Figure 3. Examples of plots of melody contours: a) extracted from MIDI file, b) hummed by QbH system user, c) found fragment of a contour, and d) plot of match error as a function of time shift

Finally, the recognition of a song is defined as its classification k to one of the database examples p:

$$k = arg\{ min(me_p) \}, \quad 1 \le p \le P \tag{5}$$

where P is the number of songs in the database.

The above approach works very well in ideal conditions of perfect note frequency and tempo reproduction in the hummed fragment. Linear scaling of the unknown melodic contour was introduced to make the algorithm more robust to possible imperfections in melody humming. Scaling in time was obtained by signal resampling with the resulting sampling period Δt_{sl} defined as:

$$\Delta t_{sl} = tc \cdot \Delta t_s, \quad tc \in \{0.8, 0.85, \ldots, 1.2\} \tag{6}$$

where:
Δt_s – original sampling period,
tc – time scaling coefficient.

The scaling coefficient takes one of nine values, from 0.8 up to 1.2. Scaling in frequency was obtained by multiplying the signal by a power of 2:

$$f_m = 2^{m/12} \cdot f, \quad m \in \{-0.5 , -0.4 , \ldots, 0.5\} \tag{7}$$

where:
f – original frequency,
f_m – multiplied frequency.

The m coefficient takes one of nine values, from -0.5 up to 0.5. The range is set to compensate pitch errors in the range of a quarter of a musical tone. After an introduction a both scaling in time and frequency a number of necessary computations increases 9x9=81 times, making the algorithm much slower, but the probability of correct melody recognition increases.

VI. System performance estimation

A series experiments were carried out on real data to evaluate the developed system's performance. The test database consisted of eight songs of popular music and four persons of different ages and genders participated in the experiment. Only one of the participants possessed musical experience. In all the cases recognition accuracy was 100%, so a distribution of the match error *me* over all the examples in the database was investigated. A matrix of match errors obtained for all the humming examples against all the examples in the database without scaling of the unknown melodic contour is shown in Table I. The lowest values are obtained for the proper examples, but other error values are not much higher. The certainty of matching is not very high, and a recognition error can occur in the case of lower recording quality. Introducing scaling of unknown melodic contour vector, both in time alone and in frequency, gave a reduction of the error level. The results are listed in Table III. The highest error reduction was observed after combining both scaling in time and frequency. In the same time the recognition certainty was increased (Table II.).

TABLE I. MATCH ERROR WITHOUT SCALING OF MELODIC CONTOUR

Match error [Hz]	**Hummed melody**							
	1.	***2.***	***3.***	***4.***	***5.***	***6.***	***7.***	***8.***
Song 1.	**47.0**	63.3	55.2	53.4	32.3	51.0	24.1	48.5
Song 2.	63.0	**55.6**	61.5	56.4	38.4	58.1	38.9	57.6
Song 3.	64.2	76.8	**47.9**	65.5	50.6	59.3	41.3	49.1
Song 4.	76.4	66.7	74.1	**32.2**	58.9	95.5	44.7	56.5
Song 5.	61.1	70.5	63.8	73.3	**12.8**	60.0	34.8	57.0
Song 6.	90.6	105.7	85.2	114	42.1	**19.1**	49.8	61.5
Song 7.	68.7	65.8	59.0	71.6	33.0	66.1	**22.0**	48.7
Song 8.	71.3	71.6	65.9	74.0	45.2	71.9	35.0	**18.3**

TABLE II. MATCH ERROR FOR SCALING OF UNKNOWN MELODIC CONTOUR BOTH IN TIME AND FREQUENCY

Match error [Hz]	**Hummed melody**							
	1.	***2.***	***3.***	***4.***	***5.***	***6.***	***7.***	***8.***
Song 1.	**39.5**	57.7	51.6	46.8	23.7	34.2	23.3	43.8
Song 2.	56.7	**31.3**	54.9	47.9	35.7	53.4	34.3	51.8
Song 3.	60.5	70.3	**29.0**	50.9	45.4	54.4	36.8	43.1
Song 4.	59	55.7	60.6	**25.6**	46.7	84.8	32.9	49.1
Song 5.	54.6	59.6	61.5	65.6	**12.3**	53.0	28.9	52.3
Song 6.	85.5	90.9	78.3	87.9	40.9	**16.8**	43.1	55.0
Song 7.	66.7	54.3	56.8	60.8	27.5	56.1	**15.5**	45.2
Song 8.	63.4	55.7	55.4	66.4	42.8	58.7	33.8	**18.1**

TABLE III. MATCH ERROR FOR DIFFERENT COMBINATIONS OF SCALING OF UNKNOWN MELODIC CONTOUR

Match error [Hz]	**Hummed melody**							
	1.	***2.***	***3.***	***4.***	***5.***	***6.***	***7.***	***8.***
No scaling	47.0	55.6	47.9	32.2	12.8	19.1	22.0	18.3
Scaling in freq.	45.0	55.1	47.7	32.1	12.3	16.8	21.3	18.1
Scaling in time	42.0	32.0	29.3	25.8	12.8	18.1	16.3	18.3
Scaling in freq. & time	39.5	31.3	29.0	25.6	12.3	16.8	15.5	18.1

Another experiment was concentrated on finding the individual capabilities of each test participant to cooperate with the proposed QbH system. The results are listed in Table IV.

TABLE IV. MATCH ERROR FOR DIFFERENT PARTICIPANTS OF THE EXPERIMENT

Match error [Hz]	**Hummed melody**							
	1.	***2.***	***3.***	***4.***	***5.***	***6.***	***7.***	***8.***
Person 1	42.1	31.3	29.0	25.6	12.3	16.9	15.5	18.1
Person 2	44.7	36.0	33.7	53.3	19.2	15.5	16.6	33.6
Person 3	40.0	35.9	34.6	24.5	22.6	24.6	16.5	19.9
Person 4	33.9	33.1	34.3	22.3	8.99	19.2	19.9	18.1

Some variation can be observed, probably due to different style of melody humming and musical ability. The differences are still not crucial to the overall system performance.

The match error for correct melody guesses in Table I. and II. seem to be very high, compared to false guesses. One of the reasons for such a situation is the lack of the human ability to rapidly change the pitch between notes in a hummed song. It produces an additive error to all matches, which does not deteriorate the algorithm's performance. This was verified on synthetic melodic contours, for which all the errors were lower, but their variability was similar, as in the real examples.

VII. CONCLUSIONS

The proposed QbH system allows for accurate melody recognition in small databases. Its performance for the small group of participants was not significantly affected by musical ability. The created application allowed to test the distribution of RMSE over the introduced examples and made it possible to investigate the effect of time and frequency scaling on the outcome of the match.

The developed system is robust to slight errors in melody humming, such as singing a few false notes, singing the melody too fast or too slow, or shifting the tune's pitch. For the testing examples the application obtained the recognition accuracy of 100%.

REFERENCES

[1] M. Casey, R.Veltkamp, M. Goto, M. Leman, C. Rhodes & M. Slaney, "Content-based music information retrieval: Current directions and future challenges," Proceedings of the IEEE, 96(4), 2008, pp. 668-696.

[2] M. Raju, B. Sundaram, P. Tansen, "A query-by-humming based music retrieval system," National Conference on Communications, NCC 2003, Madras, 2003.

[3] R. B. Dannenberg, N. Hu, "Understanding Search Performance in Query-By-Humming Systems," Fifth International Conference on Music Information Retrieval, Barcelona, 2004.

[4] S. Doraisamy, S. Ruger, "Robust Polyphonic Music Retrieval with N-grams," Journal of Intelligent Information Systems, 21(1), 2003, pp. 53-70.

[5] B. Pardo, "Finding Structure in Audio for Music Information Retrieval," IEEE Signal Processing Magazine, Vol. 23, Issue 3, May 2006, pp. 126-132.

[6] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the MUSART testbed," Journal of the American Society for Information Science and Technology, Volume 58, Issue 5, March 2007, pp. 611–762.

[7] J. Benesty, M. Sondhi, Y. Huang, "Handbook of Speech Processing," Springer, Berlin, 2008, pp. 185 – 188.

[8] P. De La Cuadra, A. Master, C. Sapp, "Efficient Pitch Detection Techniques for Interactive Music," Proceedings of ICMC 2001, International Computer Music Conference, La Habana, Cuba, September 2001.

[9] M. Dziubinski, B. Kostek, "High Accuracy and Octave Error Immune Pitch Detection Algorithms," Archives of Acoustics, No. 1, vol. 29, pp. 1 - 21, 1. 2004.

[10] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proceedings of the Institute of Phonetic Sciences 17: 97–110. University of Amsterdam, 1993.

[11] P. Boersma, D. Weenink, "Praat: A system for doing phonetics by computer," www.praat.org, 1992–2001.

Analysis of Damping Materials in a Transmission Line Loudspeaker System

Krzysztof Lusztak B.Sc.
Lodz University of Technology
Lodz, Poland
krzysiek@lusztak.pl

Michał Bujacz Ph.D.
Lodz University of Technology
Lodz, Poland
bujaczm@p.lodz.pl

Abstract — **The presented paper contains an analysis of the influence of various types of damping material in a transmission line loudspeaker enclosure on the acoustic emission spectrum. Damping of the tunnel is crucial in the design of this type of enclosures. Six types of materials traditionally used in home speaker construction were studied.**

*Keywords — **loudspeaker, vented enclosure, bass-reflex, sealed enclosure, transmission line, waveguide, damping materials, felt, wool, foam.***

I. Introduction

The most common loudspeaker enclosure types are either closed [1] or vented (bass-reflex type) [2,3]. These two types of enclosures can be relatively easily and accurately modeled, allowing mathematical simulations to be used in efficient design Moreover, because of the small number of variables, the enclosures can also be quickly tested through trial and error with very good effect. For the sealed enclosure only the volume of the loudspeaker and the amount and type of damping materials can be changed, while the bass reflex enclosure (utilizing the Helmholtz resonator) introduces two more variables - the length and the diameter of the resonator (which are in fact correlated [3]). Thanks to these features the closed and bass-reflex enclosures have dominated the market and nearly forced out all other enclosures types from commercial domestic applications. Due to the stretching of the speaker response toward the low frequencies with relatively low design effort, the bass reflex enclosure are the predominant type used in home audio. The trend is so strong that many transducers are designed specifically for use in bass reflex enclosures.

However, these types of enclosures have several disadvantages. The resonant frequency of a loudspeaker mounted in a sealed enclosure increases, but at the same time begins to fall earlier near the low frequencies, which may result in the listener perceiving a lack of low-frequency sound material. The bass reflex type enclosure introduces a delay into the played sound – due to the fact that within the range of frequencies near the tuning frequency of the resonator it is the resonator itself (not the loudspeaker) that is the source of the sound. The bass reflex enclosure also increases the risk of overdriving the speaker at frequencies below tuning, especially if it is poorly designed.

There is a loudspeaker enclosure type that avoids the aforementioned problems – the transmission line enclosure [4,5]. This type of enclosure contains a tunnel behind the loudspeaker, with specific parameters serving as a waveguide. The resonance phenomena doesn't appear in transmission line enclosure, so there is no audible lag in the music material [6]. The response of the speaker has a natural decrease of 12dB/octave in the direction of the low frequencies and this decline begins later than in closed speaker casings. The waveguide is designed to use the rear surface of the speaker membrane, reverse it in phase and use as a second source of sound. Unfortunately, this type of enclosure is difficult to design. The number of variables is very large: the length of the tunnel, the location of the speaker relative to the entry of the tunnel, the tunnel's capacity, the cross sectional area of the beginning and end of the tunnel, and their relationship to each other, the type and location of the damping material, the number of turns in the tunnel. Most of these variables can be simulated by using appropriate modeling software. Unfortunately it is impossible to properly simulate the placement and the type of damping material, which is crucial in this type of construction. To understand why the damping is crucial it is necessary to first have a good grasp of the principle of operation of a transmission line enclosure.

II. Theory behind transmission line enclosures.

The basic principle of a transmission line enclosure is that the wave emitted from the front of the speaker membrane is combined with the wave from the rear of the membrane after it has traveled the transmission line tunnel and has shifted in phase. The shift depends on the length of the tunnel and the wavelength of the sound. To reverse the phase by 180 degrees (to provide the same phase of the emission front side of the membrane and end of the tunnel) at a frequency of 30Hz the tunnel would need to be half as long as the wavelength of that frequency, i.e. approximately 5.7m (assuming sound speed of 344m/s). In practice, such a long tunnel is not necessary. The vector arithmetic shows that even when the phase is shifted by about 90° (with the length of the tunnel equal to ¼ wavelength) the summary emission is greater than the emission of only the front side of the membrane (Figure 1). So for a "resonance" at 30Hz the tunnel length only needs to be about 2.9m.

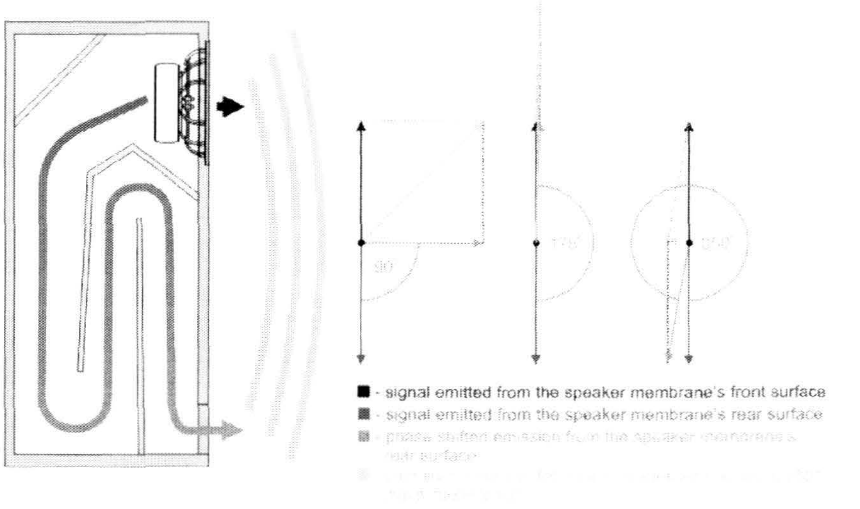

90
180
270
■ - signal emitted from the speaker membrane's front surface
■ - signal emitted from the speaker membrane's rear surface
■ - phase shifted emission from the speaker membrane's rear surface

Figure 1. Phase phenomena in a transmission line loudspeaker system.

The maximum combined emission (from speaker and the tunnel) occurs when the phase is shifted close to 180° , which will accentuate this range of frequency characteristics. For the 2.9m tunnel the first "resonance" would occur for frequencies near 60Hz. For frequencies near 120Hz the phase shift will be close to 360°, which will mean the wave emitted through the end of the tunnel is in opposite phase to that from the speaker, and as a result we will see a collapse on the combined emission frequency characteristics (the first "anti-resonance"). Phenomenon like this of un-damped tunnel will occur periodically throughout the higher frequencies as illustrated in Figure 2.

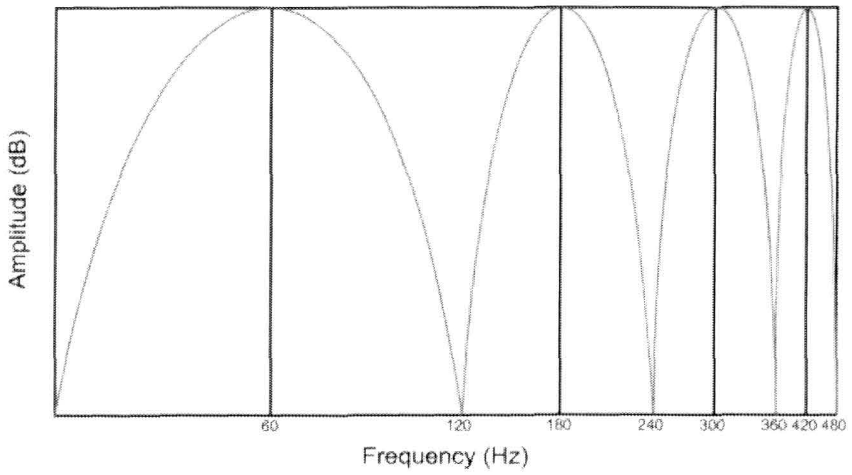


Figure 2. Ideal combined emission from un-damped transmission line loudspeaker system (tunnel length 2.9 m).

To minimize the resulting unevenness of the characteristics, while keeping a strong amplification at low frequencies, the emission of high frequencies needs to be attenuated. To accomplish this, damping materials lining the tunnel are used. The main task for the damping material is to limit tunnel's first "anti-resonance", which is the most complicated task when designing transmission line enclosures [7].

III. Analysis of Damping Material

The most commonly used damping materials in DIY audio are various foams, felts and upholstery wools. The presented study compared the following materials:

- three layers of technical felt with a thickness of 3 mm each (Figure 3.1),
- upholstery wool with a thickness of 3 cm (Figure 3.2),
- plain foam with a thickness of 2 cm (Figure 3.3),
- plain foam with a thickness of 4 cm (Figure 3.4),
- pyramidal foam with a base thickness of 1 cm and a total thickness of 4 cm (Figure 3.5),
- wave-profiled foam with a base thickness of 2 cm and a total thickness of 4 cm (Figure 3.6).

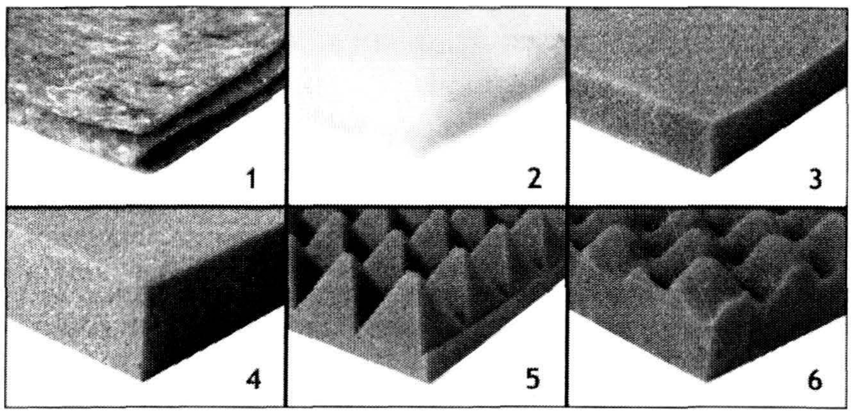


Figure 3. Photos of the damping materials used in tests.

Parameters of the tested enclosure:

- tunnel length: 1.8 m,
- area of the beginning of the tunnel: 2 x Sd of the speaker,
- area of the end of the tunnel: 0.8 x Sd of the speaker,

speaker placed at the beginning of the tunnel.

Speaker used in measurements: SEAS L22 RN4X/P.

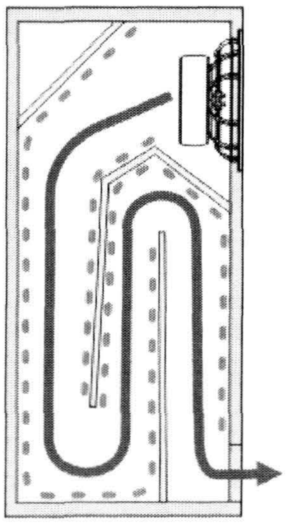


Figure 4. Diagram of the test enclosure, along with position of damping material (dotted lines).

Each type of damping material was mounted in the same places marked in Figure 4. For each type of material three measurements were performed:

- measurement in the far field (1 m from the enclosure) of the response as a function of frequency - the measurement shows combined emission of the tandem speaker + enclosure – top line in Figures 6 to 12,
- measurement in the near field (1 cm from the exit of the tunnel) of the response of the tunnel as a function of frequency - measurement of the tunnel shows a range of tunnel emission – middle line in Figures 6 to 12,
- measurement of the speaker impedance - this measurement shows the tunnel's self-resonance, depends on its length – bottom line in Figures 6 to 12.

Measurements were made in an acoustic chamber with a 1 m long microphone (eliminating the impact of the sound wave reflections from the tripod) with a WM61A Panasonic microphone cartridge. A laptop with an M-Audio Audiophile USB sound card and freeware software Audua Speaker Workshop [8] was used for data collection and analysis.

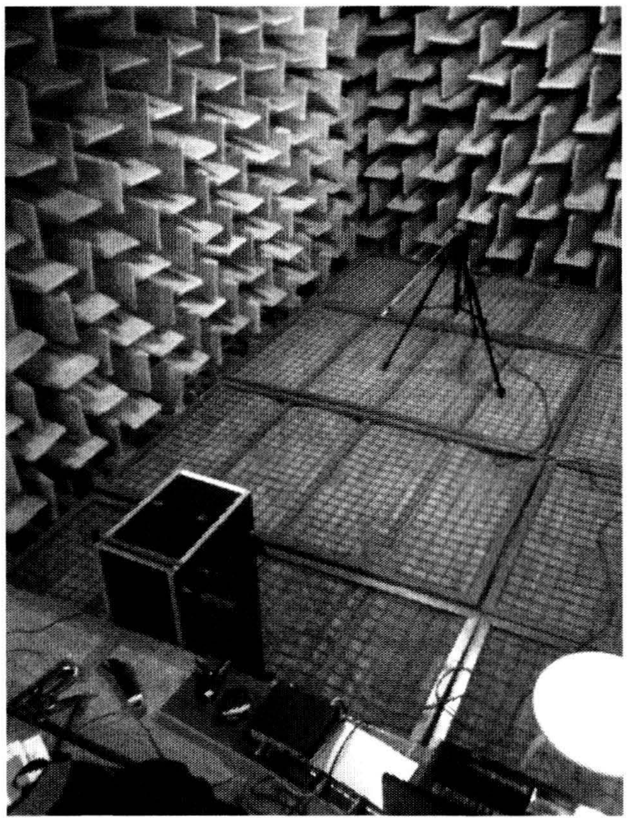


Figure 5. Measurement setup in an anechoic chamber.

IV. MEASUREMENT RESULTS

The measurement results are presented in the same scale (10 dB grid), the frequency is limited to 5 kHz. Measurements of the tunnel response (collected in the near field, 1 cm from the exit of the tunnel) are shifted by -45 dB to better show the measured dependences.

- control measurements performed without any damping material:

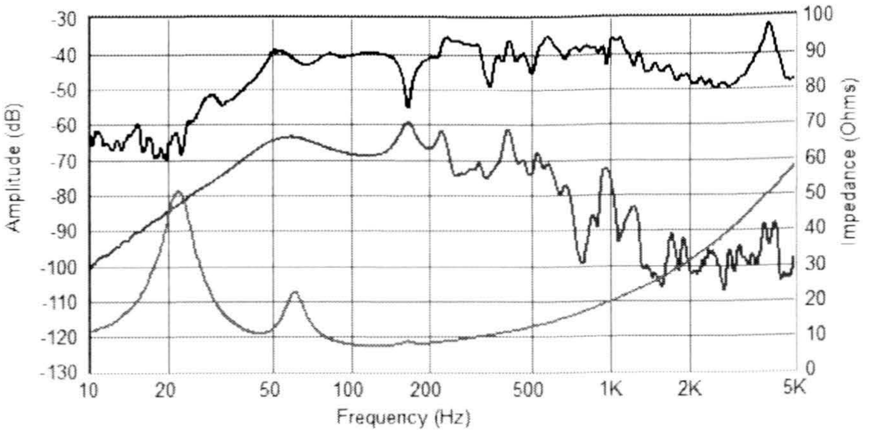


Figure 6. Measurements of un-damped enclosure. Far field (top), near field at tunnel exit (middle), speaker impedance (bottom).

It is clear that the anti-resonance occurring at about 170Hz is very strong. The emission from the tunnel at this frequency is also very strong. The tunnel self-resonance is about 44Hz, which in the classical formula for resonance

$$\frac{344\,\text{m/s}}{44\,\text{Hz}} \cdot \frac{1}{4} = 1{,}95\,\text{m} \qquad (1)$$

suggests that the tunnel length is 1,95 m. The physical length of the tunnel is only 1.8 m. It can be concluded that shape of the tunnel with multiple turns decreases the average wave velocity, because of that the speaker performs as if connected to a longer tunnel. It is expected that with the use of damping material the velocity reduction will be even greater.

- three layers of technical felt with a thickness of 3 mm each:

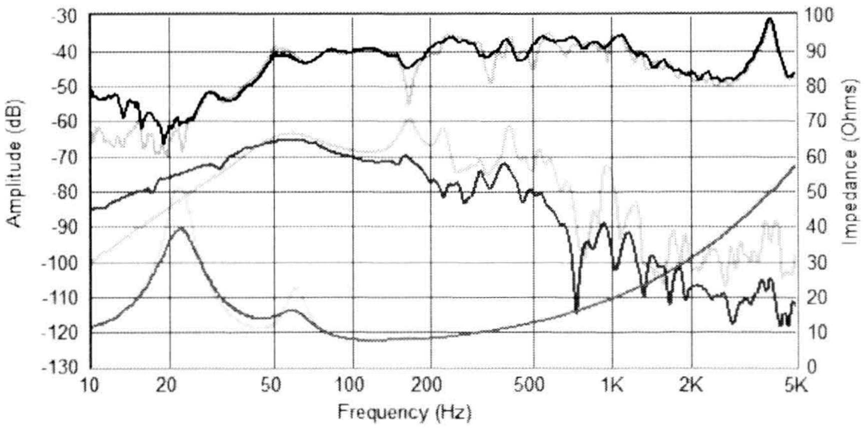


Figure 7. Measurements of enclosure damped with technical felt. Far field (top), near field at tunnel exit (middle, shifted by -45 dB), speaker impedance (bottom), and control measurement without damping (background).

It can be noticed the reduction of emissions from the tunnel in the higher frequencies, the impact of the tunnel on the combined (far field) emission characteristics is much smaller. Unfortunately, the antiresonance at around 170Hz is still clearly visible and the combined characteristics is far from smooth.

- upholstery wool with a thickness of 3 cm:

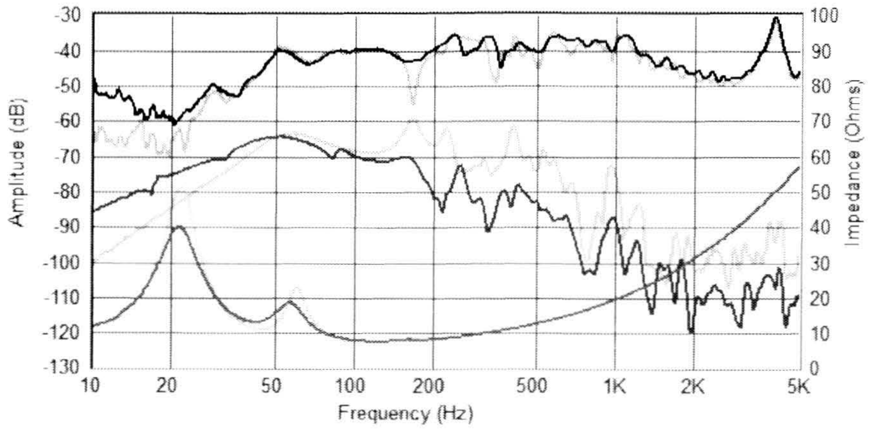


Figure 8. Measurements of enclosure damped with upholstery wool. Far field (top), near field at tunnel exit (middle, shifted by -45 dB), speaker impedance (bottom), and control measurement without damping (background).

The effect of the upholstery wool is similar to that caused by the felt. The antiresonance decreased slightly, but still the far field characteristic is very uneven. The tunnel's self-resonance is reduced to about 42Hz, which suggests a greater reduction of the wave velocity in the tunnel.

- plain foam with a thickness of 2 cm:

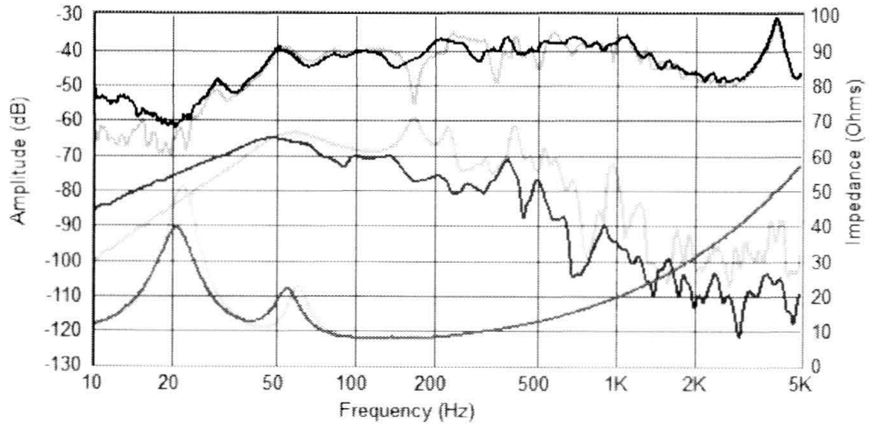


Figure 9. Measurements of enclosure damped with plain foam 2cm. Far field (top), near field at tunnel exit (middle, shifted by -45 dB), speaker impedance (bottom), and control measurement without damping (background).

The characteristics are similar to those measured for the upholstery wool and felt. The antiresonance is still clearly marked, but has changed its position on the frequency axis. This is due to the significantly decreased self-resonance frequency of the tunnel to about 39Hz, and thus even larger wave velocity reduction.

- plain foam with a thickness of 4 cm:

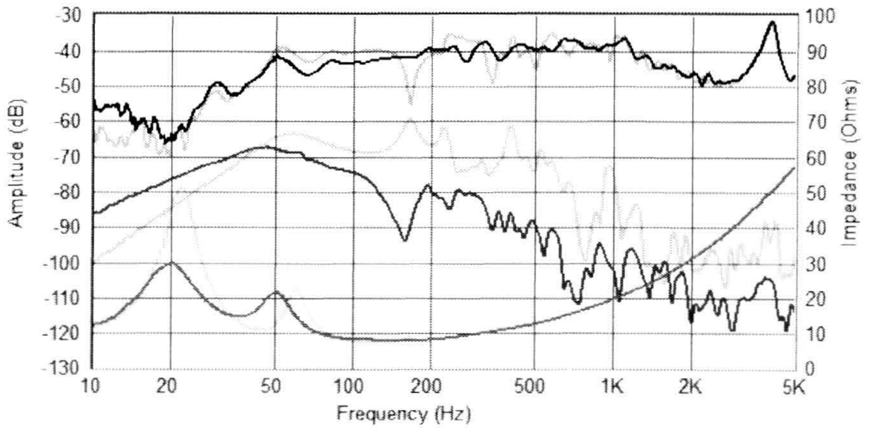


Figure 10. Measurements of enclosure damped with plain foam 4cm. Far field (top), near field at tunnel exit (middle, shifted by -45 dB), speaker impedance (bottom), and control measurement without damping (background).

The effect is significantly different from the first three materials. The antiresonance is practically unnoticeable. The emissions from the tunnel are very limited in the frequency range 100-200Hz. The self-resonance of the tunnel is located at about 36 Hz. The use of this material created the effect of a virtual lengthening of the tunnel to 2.4 m. Unfortunately, the expense is much weaker emission of the tunnel in the lower frequency range, which get less profit from the work of the tunnel.

- pyramidal formed foam with a base thickness of 1 cm and a total thickness of 4 cm:

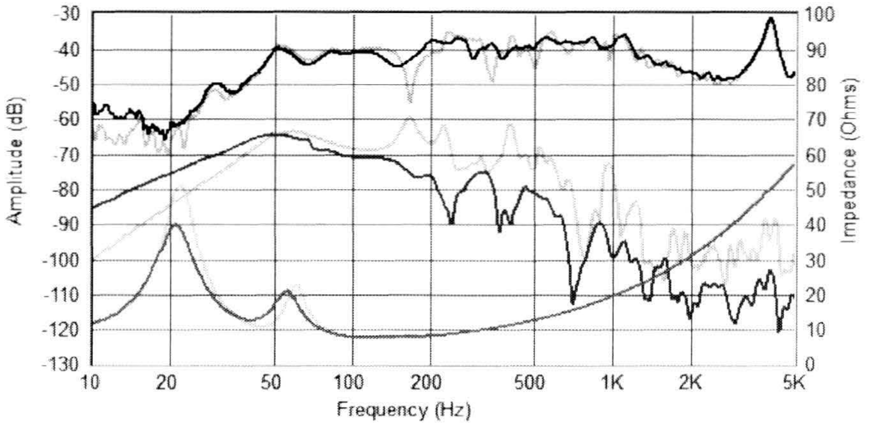


Figure 11. Measurements of enclosure damped with pyramidal foam. Far field (top), near field at tunnel exit (middle, shifted by -45 dB), speaker impedance (bottom), and control measurement without damping (background).

Pyramidal formed foam with a base thickness of 1 cm and a total thickness of 4 cm works almost exactly like the plain foam with a thickness of 2 cm, leading to the conclusion that the shape of the padding material is less significant than its volume and density.

- wave-profiled foam with a base thickness of 2 cm and a total thickness of 4 cm:

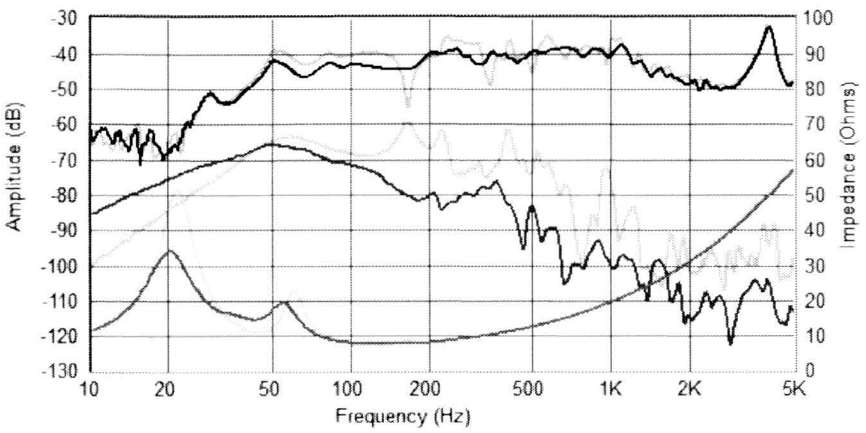


Figure 12. Measurements of enclosure damped with wave-profiled Far field (top), near field at tunnel exit (middle, shifted by -45 dB), speaker impedance (bottom), and control measurement without damping (background).

The wave-profiled foam's performance is something between the plain foam with a thickness of 2 cm and the plain foam with a thickness of 4 cm. Some antiresonance can be seen, but the characteristic is relatively even.

V. Conlusions

On the basis of presented measurements the following conclusions can be drawn:

Foam is able to significantly slow down the wave velocity in the tunnel, so it is possible to virtually lengthen the tunnel. This may be utilized to physically reduce the size of the enclosures.

The profiling of the foam surface is not important. What matters is the amount of foam volume in the tunnel,

It is fairly easy to reduce the emission of the tunnel (felt, upholstery wool), but to obtain significant reduction and characteristic smoothing is much more difficult and requires to use other, thicker materials,

Foam as the damping material is overall the most advantageous,

Further studies can focus on attempts to model the influence of the foam thickness on the emission characteristics to find the optimal amount of padding. It can also be speculated that a combination of different damping materials such as foam and felt or upholstery wool can give a satisfactory result.

VI. References

[1] R. H. Small, "Closed-box loudspeaker systems, parts I and II," J. Audio Eng. Soc., vol. 20, pp. 798-808, Dec. 1972; vol. 21, pp. 11-18, Jan./Feb. 1973.

[2] Thiele, A. N., "Loudspeakers in Vented Boxes: Parts I and II," J. Audio Engineering Soc., Vol 19, No. 5, May 1971, pp 382-392 (Reprinted from a 1961 publication in Proc. IRE Australia)

[3] R. H. Small, "Vented-box loudspeaker systems, parts I-IV," J. Audio Eng. Soc., vol. 21, pp. 363-372, June 1973; pp. 438-444, July/Aug. 1973; pp. 549-554, Sept. 1973; pp. 635-639, Oct. 1973.

[4] A. R. Bailey, "The transmission line loudspeaker enclosure," Wireless World, May 1972.

[5] B. Olney, "The acoustic labyrinth", Electronics, April 1937.

[6] A. R. Bailey, "A non-resonant loudspeaker enclosure," Wireless World, October 1965.

[7] G. L. Augspurger, "Loudspeakers on damped pipes," J. Audio Eng. Soc., vol. 48, pp. 424-436, May 2000.

[8] http://www.speakerworkshop.com

SESSION 4:
HUMAN-COMPUTER INTERACTION

Design of a Generic Head-Mounted Gaze Tracker for Human-Computer Interaction

Darius Mazeika
Kaunas University of technology
Mechatronic Department
Kestucio street 27, 44312 Kaunas, Lithuania
da.mazeika@stud.ktu.lt

Andrea Carbone, Edwige E. Pissaloux
Université Pierre et Marie Curie (Paris 6), ISIR
4 place Jussieu
750005 Paris, France
Edwige.Pissaloux@upmc.fr

***Abstract* — This paper proposes a design of a generic head mounted gaze tracker for both, infra-red and visible, spectra. The system design takes into account inter person morphological and visual perception parameters. The proposed low cost gaze tracker is optimized in weight. An algorithm for eye robust tracking in the presence of different type of eyes and eye movements is presented. The gaze tracker potential usage as an assistance of PC interactions of the upper limb motor impaired is outlined.**

***Keywords** — gaze tracker, visible spectrum, assistance of upper limb impaired interaction*

I. Gaze tracker current STATUS and Future applications

Eye movements are a key element for various cognitive studies such as human attention, interaction (human, robots, virtual environments), navigation, reading, painting analysis, etc., and, more recently, for the design of the new environments named ARE (Attention Responsive Environment, [1]). The usage of the gaze tracker as a tool for such activities requires the acquisition of specific skills for control of eye movements (dwelling, displacement in precise direction and at the appropriate speed, eyelids blinking at specific time, etc.).

Gaze tracker is a system which allows not only to track the eye movements over a 2D surface located at a fixed distance (such as a computer screen, for example), but it allows also to identify a 3D gazed point (point of gaze) whose distance to the observer can vary. A head-mounted gaze tracker is most suitable for interactions at different depths. It can have one or two eye cameras which film one or two eyes, and one camera for scene image acquisition.

However, all the existing gaze trackers ([2]) are rather expensive and based mainly of infra-red (IR) technology only. The IR technology facilitates the image processing, but there is a lack of epidemiological data on the usage of the IR technology during several hours per day (what is usual case in PC gaming or internet usage) [3].

Therefore, a design of a low cost gaze-tracker adapted for short and long time usages is an objective of the AsTeRICS, FP7 ICT project [4]. This paper addresses the design of such gaze tracker. Section 2 briefly introduces the gaze tracker specifications. Section 3 provides main details of an adaptive mechanical support for the gaze tracker architectures. Section 4 shortly presents the basic software for visible spectrum eye tracking, while Section 5 outlines the designed gaze tracker possible exploitations as assistance for PC screen interaction for the upper limb impaired people. Finally, Section 6 proposes the current status of the gaze tracker design and its future improvements.

II. SPECIFICATIONS of a Head-Mounted gaze tracker

A generic gaze tracker has to satisfy three classes of parameters: intra-person (or human) parameters, generic design parameters and parameters of optimal realization.

Human parameters are head sizes, distances "eye-camera(s)", allowed head movements, eye illumination (infra-red (IR) or visible).

Generic design parameters are vision system configuration (mono/binocular system); adaptability to scene illumination; precision of gaze detection adaptable to targeted application ; interchangeable parts (for system different configurations); possibility to add additional sensors (such as an inertial sensor for example); ease to wear; ease to use.

Parameters of an optimized (or optimal) realization are the following: minimized cost, minimized weight; fast realization time; reduction of the obtrusiveness of the field of view.

The above listed parameters are implemented in different gaze tracker architectures.

Typical gaze tracker architecture has therefore, 3 synergetic components:

- mechatronic support, i.e. system framework and support for sensors and control subsystems;
- sensors for images management (acquisition and stabilisation);
- specific algorithms (software) for image processing.

The mechanical (& control) support of the targeted head-mounted gaze-tracker has to respect the following design criteria:

- have the maximal degrees of freedom (DOF) which allow to adapt it to the human specific needs (inter-personal variations) such as anatomy of the head, vision capability, human daily activities, and their temporal evolution;
- have a good mechanical resistance,
- be lightweight,
- be easy wearable,
- have a sufficient temporal power autonomy,
- be of a low price for manufacture,
- be built with few off-shelf standard components.

Two types of sensors are used: cameras (visible and IR for eye images, and visible camera for scene images) and inertial measure unit (IMU).

The camera should be exchangeable (of a visible and IR spectrum), and should acquire images of the best quality (with a minimal noise (ideally, without noise) and of high contrast. Eye images should be in centre of acquired eye images. The field of view of scene camera should subtended by a solid angle of at least 60°. Cameras should be lightweight and of a small size. The IR illumination light should have an adaptable power (selection of the right IR light wavelength, selection of the correct current, selection of the right LED elements angle and units).

The IMU should provide high precise results, should be easy to calibrate, of a low price and low power consumption; furthermore, it should be lightweight and of small size.

The algorithms should work with stabilised images of a very good resolution. They should provide a simple calibration, fast, precise and reliable eye tracking, and fast matching between eye and gaze positions. All processing should be optimized in space and time for their fast processing by a processor of an embedded system.

III. ADAPTIVE GENERIC MECHANICAL SUPPORT FOR A GAZE TRACKER

The mechanical support for a generic adaptive gaze tracker is built upon a standard helmet (used by welders), to which different original specific elements are added. The number, types and parameters pertinent for gaze tracker final usages define the degrees of freedom (DOF) of gaze tracker; they are also named adaptability parameters.

Two architectures of the head-mounted gaze tracker are considered: a direct gaze-tracker with (IR or visible spectrum) camera (Figure 1 a), and an indirect gaze-tracker with IR camera (Figure 1b).

In a direct gaze tracker, the eye camera(s) directly films (two) eye(s) , while the forehead camera films the observed scene in visible spectrum. In the case of visible spectrum, camera takes color images of the human eye while in the case of IR setup, camera takes IR images of a human eye. In both cases, the variations of 3D mechanical parameters of the boom arm supporting the camera is directly linked to head morphology, perception capability of the human eye and acquisition parameters of the camera (field of view, resolution, focal distance; eye sensitivity to IR illumination in the case of IR camera).

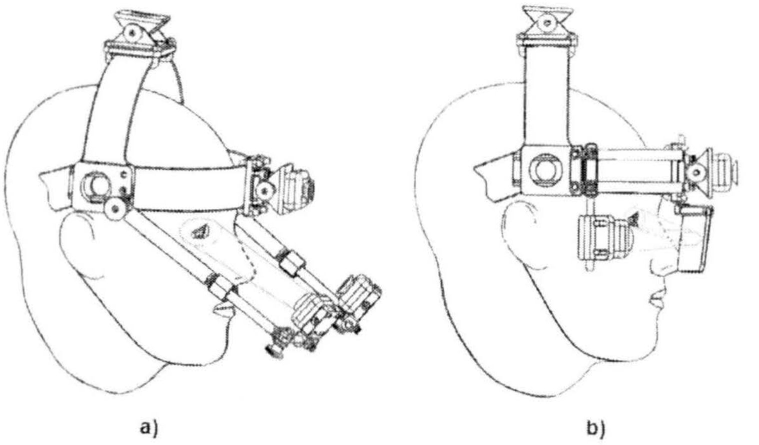


Figure 1. Head-mounted Gaze-tracker architectures: a) 2-camera direct system, b) 1-camera indirect system.

In an indirect camera system, the IR camera, placed on a side of a face (close to the ears), films the image of an eye reflected by a hot mirror. The hot mirror is transparent to visible light, and it reflects non-visible IR light to IR camera(s). The angle between the hot mirror and IR camera determine the quantity of the reflected light thus the quality of the IR image of the eye. IR light location can be in one of the two settings: co-planar with the IR camera or not (usually, located on a specific IR light arms (Figure 2 right).

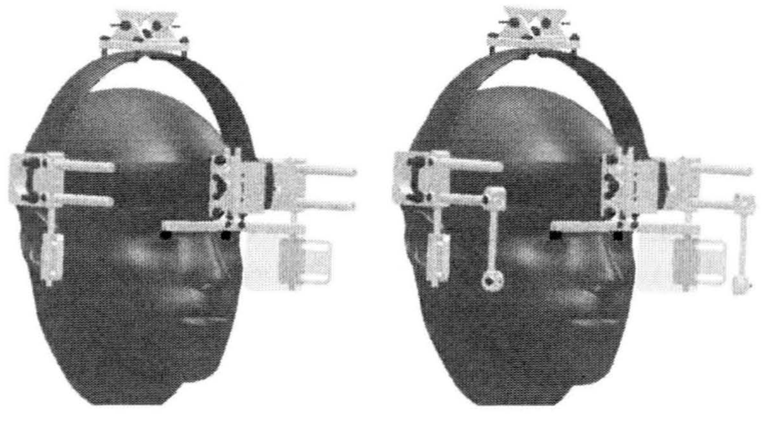


Figure 2. Indirect IR gaze tracker without (left) and with the arm for IR illumination source.

3.1. Parameters of direct gaze tracker.

The mechanical support for direct gaze tracker with one boom arm has 10 DOF (cf. figure 3) ; there are :

- 4 DOF of the camera telescopic boom arm ;
  ($\alpha = 0 \div 75°$; $\beta = 0 \div 360°$; $\gamma = 0 \div 200°$ *and* $S_1 = 120 \div 200mm$);
- 3 DOF of support for the scene camera;
  ($\delta = 0 \div 360°$; $\varepsilon = 0 \div 360°$ *and* $S_2 = 10 \div 30mm$);
- 1 DOF of place for the scene camera;
  ($\zeta = 0 \div 360°$);
- 2 DOF of support for gaze tracker basic control;
  ($\eta = 0 \div 30°$ *and* $\theta = 0 \div 30°$).

The DOF/parameters of boom arm play the following roles (their 3D position is usually defined with respect to human body main plans: frontal, sagittal, transverse) :

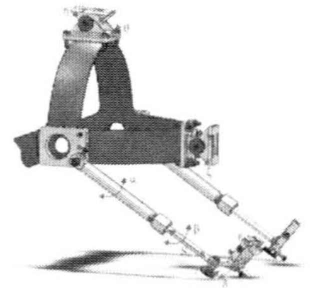


Figure 3. 17 DOF of a direct gaze-tracker.

-the angle α helps to set the camera position in front of the eye, in order to acquire the eye images with eye located in images centre;
- the angle β adapts the camera rough position to human face, in such a way that the image of the eye will be parallel by the transverse plane of the body/face;
- the angle γ tunes the camera position to human face (the camera is adjusted to parallel (left or right) lateral and median body planes;
- the displacement S_1 allows to tune the length between the boom arm ground element and boom arm rod element. This is needed to acquire the eye images of good quality; this parameter is tightly coupled to camera resolution and its optics (focal, field of view).

The support for eye camera has 3 DOF: two angles δ, ε and displacement S_2 :
- the angle δ keeps the camera position parallel to the body/face frontal plane (the eye camera should be in vertical position when boom arm ground element *angle α* is changed);
- the angle ε adapts the camera position to human eye parallel to the transverse plane when boom arm rod element *angle γ* is changed;
- the displacement S_2 adjusts the length between boom arm rod element and camera mounting plate in order to acquire the eye images parallel to the transverse plane.

The support for the scene cameras attached to the helmet has 1 DOF, angle ζ ; this angle keeps the scene camera position in environment (parallel to the transverse plane).

The support for gaze tracker basic control system (on the top of helmet) has 2 DOF defined by angles: η and θ. *The angle η* adapts the box of the electronics (control and IMU) in the position which must be in parallel with transverse plane. *The angle θ* keeps the gaze tracker control unit position (IMU included) in parallel with respect to the transverse plane.

The proposed direct eye tracker concept has 17 DOF, therefore, it is adaptable to several inter personnel morphologies and different conditions of interaction. Figure 4 shows a front view of the realised direct gaze tracker for one eye tracking. The white parts – boom arm and support for scene camera- are original (designed with a rapid prototyping tool, RTP).

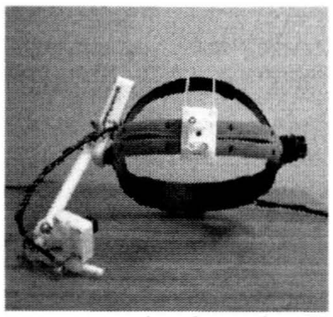


Figure 4. Direct gaze tracker front view (1-eye system)

3.2. Parameters of indirect gaze tracker.

Figure 5 shows the model of the indirect gaze tracker. This prototype of gaze tracker has 15 DOF with two hot mirrors (cf. figure 5), and only 9 DOF with one hot mirror. The design parameters are :
- 1 DOF of IR camera arm (S_1 = *0÷60mm*);
- 2 DOF of side camera mounting plate;

(α = *0÷200° and* S_2 = *0÷35mm*);
- 3 DOF of hot mirror arms;

(β = *0÷90°;* S_3 = *0÷25mm and* ζ = *0÷90°*);
- 1 DOF of place for the scene camera (γ = *0÷30°*);
- 2 DOF of support for gaze tracker basic control;

(δ = *0÷30° and* ε = *0÷30°*).

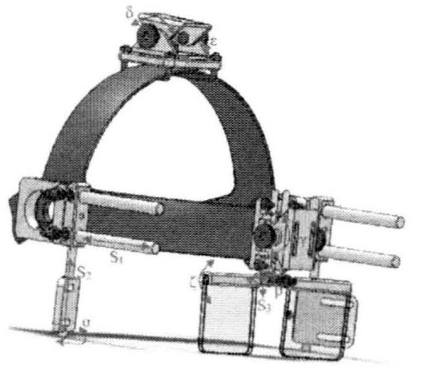


Figure 5. 9 DOF of an indirect gaze-tracker.

The IR camera arm has 1 DOF, the displacement S_1. S_1 adjusts the length between helmet camera/IR light ground plate and arm what is important for acquisition of the eye images of good quality. The distance variation is tightly related to camera (resolution) and its optics (focal, field of view).

The side camera mounting plate has 2 DOF : the angle α and displacement S_2. The angle α adapts the camera position to hot mirror when IR camera arm *displacement S_1* is changed. *The displacement S_2* tunes the camera's position to reflected eye on the hot mirror when hot mirror arm *displacement S_3* is changed.

The hot mirror arm has 3 DOF, there are two angles β, ζ and displacement S_3. The angle β adapts the hot mirror position to side camera when displacement S1 changes. The displacement S3 adjusts the length between camera/hot mirror ground plate and hot mirror catcher element in order to acquire the eye pupil images of good quality and enough reflecting image. The angle ζ adapts the hot mirror position to side camera in order to obtain the reflected human eye pupil at the centre on hot mirror in the frontal plane. The place for the scene cameras has 1 DOF : angle γ which adapts the scene camera position to environment.

The support for gaze tracker power supply and IMU has 2 DOF: angles δ and ε. The angle δ keeps the system control in parallel to the attached inertial measure unit (IMU), while the angle ε keeps the gaze tracker control unit position (IMU included) in parallel with the transverse plane.

Figure 6 shows the designed gaze-tracker system adaptable to inter-person anatomy variations. All design parameters target the acquisition images of high quality where features pertinent to image and vision processing will be easier, faster and more reliably detected and processed. This gaze tracker takes into account the comfort and medical security of IR technology use.

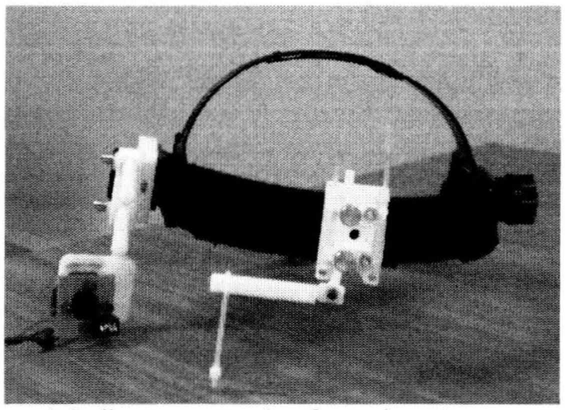


Figure 6. Indirect gaze tracker front view (1-eye system).

3.3. Design optimization.

A wearable gaze-tracker must be easy to wear and lightweight. As said before every original component will be printed with a rapid prototyping tool, RPT, (3D printer), so it is possible to minimize the system weight by removing some plastic material inside of designed specific components. The minimisation is a trade-off between component weight and its mechanical resistance. Practically, if the optimized components are done by CNC machining the weight optimization is not possible. However, it should be noticed that the weight minimization will lead to component of higher price.

In the RPT, when lightened components are printed, the 3D printer fills-in the empty places with a material (usually with the ABS BASS). The ABS BASS mechanical resistance is less compared to the ABS as the material is porous (instead of solid one). Figure 7 gives an example of weight optimisation: the blue parts have been replaced by the ABS BASS. The weight optimisation of all original parts of the adaptable gaze tracker reduces its initial weight by around 15%.

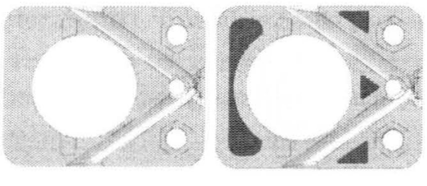


Figure 7. Section view of "Helmet camera/IR light ground plates". a) non-lightened, b) lightened

IV. EYE TRACKING ALGORITHM FOR HM-GAZE TRACKER.

The whole gaze tracking process for HM-system includes several steps. All of them target to recover a 3D point from 3 images acquired with the gaze tracker. Here after, vision eye detection and tracking approach is briefly outlined.

Probabilistic approaches to eye tracking seem better track the eye in close-up images acquired with a low-cost vision "only" camera and with uncontrolled illumination conditions.

The defined probabilistic approach [5] combines two concepts: the sequential Monte Carlo algorithms (SMC, known also as a particle filter) and the radial symmetry transform.

The SMC algorithm allows formulating multi-hypothesis in order to explore the state space (all probable positions of the eye in the next image) using the currently acquired image in order to estimate the position of the eye in the next image. As a particle filter converges to the true posterior probability density function (pdf) with the increase of particles' number (theoretically, with their infinite number), the SMC is time consuming solution space exploration method.

The radial symmetry [6] guides the potential particles' selection and therefore improves the temporal performances of the particle filter. The radial symmetry has been selected because of eye symmetric shape (the iris can be modelled with an ellipse of center (c_x, c_y)) and a potential eye movement in any direction from the current pixel p = {x, y}. This transform accumulates contributions of magnitudes and orientations of luminosity function of pixels in the p neighbourhood in different distances (radii) r from p in the gradient orientation.

Figure 8 outlines the proposed SMC-radial symmetry approach.

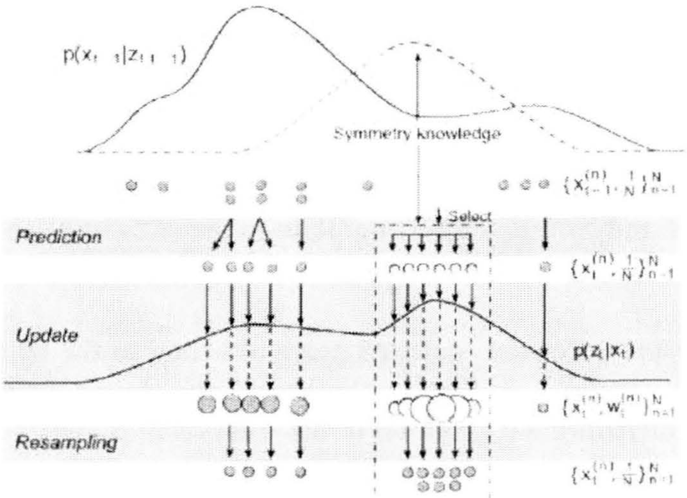

$p(x_{t-1}|z_{t-1})$
Symmetry knowledge
Select
$\{x_{t-1}^{(n)}, \frac{1}{N}\}_{n=1}^{N}$
Prediction
$\{x_t^{(n)}, \frac{1}{N}\}_{n=1}^{N}$
Update
$p(z_t|x_t)$
$\{x_t^{(n)}, w_t^{(n)}\}_{n=1}^{N}$
Resampling
$\{x_t^{(n)}, \frac{1}{N}\}_{n=1}^{N}$

Figure 8. Radial symmetry guided particle filter (the grey particles x_t are generated at instance t according to the probability $p(x_t/x_{t-1})$, while white particles are propagated thank to the system status z_t. $q_{obs}(x_t/x_{t-1}, z_t)$.

The particles' selection dynamic model is formulated as a Gaussian mixture including observation at time step t given by a radial symmetry detector. Whenever the symmetry knowledge rises above the known pdf, the old set of samples is replaced by new set of samples such that sample density better reflects posterior pdf. This eliminates particles with low weights and selects (or generates) particles in regions of highest probability for eye detection.

Consequently, the radial symmetry
a) robustifies iris tracking via a particle filter as it generates only the correctly predicted next positions of the eye,
b) reduces the volume of calculation,
c) handles abrupt motion and
d) automatically recovers from track loss (due to eyelids occlusion for example).

V. Gaze tracker for Assistance of Upper Limb Motor impaired.

One of the targeted applications of a low-cost gaze tracker is the assistance of screen interactions of the upper limb impaired.

The basic scenario targets the simulation of a PC mouse operations via the gaze what includes mouse displacement and mouse clicking. The mouse displacement toward an object is a convenient following the eye movements from the start to the end points. The mouse clicking operation (a selection of an object) is defined via a dwell time.

It should be noticed that other screen local and global operations can be implemented via face expression recognition (and processed using ASM, active shape models).

More classic application of the gaze tracker is its usage in order to control the environment, such as a door opening, room light switching on/off or music player control. The environmental control is usually performed via a PC screen displayed specific grid, where different control options are represented by icons (cf. Figure 9 for an example of such interface used in the frame of the AsTeRICS project).

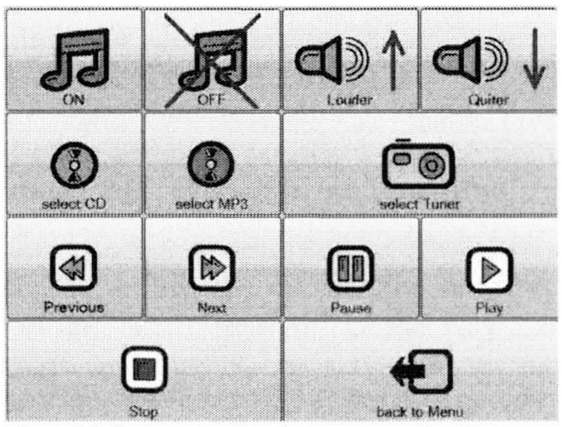


Figure 9. A grid for gaze tracker control of a remote HIFI.

In technologically more elaborated future scenario, the PC mediation will be avoided as objects of the attention responsive technology environment will be able to directly interact with a gaze in order to select the appropriate operation for execution. In such case a dedicated software continuously tracks the end-user eye(s) and gaze point in order to determine where in the environment eyes are looking

VI. Conclusion.

This paper proposes the design of a mechanical support for a generic low-cost gaze tracker. Its originality resides in the system adaptability to human inter-personal data, to sensors characteristics and data acquisition conditions via multiple degrees of freedom (DOF) of the designed system. Two generic implementations of a gaze tracker have been presented: direct and indirect.

Through its numerous DOFs, the direct gaze-tracker can adapt to a wide selection of low cost CCD cameras and associated optics, and to large variations of morphological and perceptual capability of the end-users.

Through its numerous DOFs, the indirect gaze-tracker can adapt to a wide selection of low cost IR cameras, wide selection of hot mirror sizes and shapes, wide selection of the associated side camera optics and to various capability (morphological/perceptual) of the end-users.

The direct gaze-tracker configuration can be used with any, visible or IR, spectrum.

The proposed adaptable systems and associated software will allow to gather gaze scan paths for detailed human eyes behavior registration when performing different cognitive tasks (such as navigation, reading, human-human interaction, human-robot interaction, etc.). The design and implementation of such software is one of the future directions on the AsTeRICS project. Once application software and hardware integrated, the whole system will be evaluated with primary and secondary users in three European countries (Poland, Spain, Austria) for its improvements and quality life estimation studies.

Acknowledgment

This research project is supported by the EU FP7 ICT project "AsTeRICS". The authors thank all students from the UPMC and MIT (USA) for their contributions to this work.

References

[1] Bonino, D., Castellina, E., Corno, F., Gale, A., Garbo, A., Purdy, K., Shi, F. : A Bleuprint for Integrated eye-controlled environments, Univ Access Inf Soc (2009) 8:311–321, Springer

[2] http://www.cogain.org/wiki/Eye_Trackers

[3] http://www.cogain.org/w/images/c/c8/COGAIN-D5.4.pdf , 2008

[4] http://www.asterics.eu/

[5] Martinez, F., Carbone, A., Pissaloux, E., Radial Symmetry guided Particle Filter for Robust Iris Tracking, Proc. CAIP 2011, Spain, pp. II-531-540

[6] Loy, G., Zelinsky, A., Fast radial symmetry for detecting points of interest, IEEE PAMI, 25 (2003), 959-973

A Preliminary Study on Passive Gaze Tracking System for HCI

Jacek Rondio
Institute of Electronics
Technical University of Łódź
Łódź, Poland
jacek.rondio@dokt.p.lodz.pl

Paweł Strumiłło
Institute of Electronics
Technical University of Łódź
Łódź, Poland
pawel.strumillo@p.lodz.pl

***Abstract* — The article presents a seminal results of a work on short review of current eye gaze tracking algorithms and systems, focusing mainly on the passive, non-intrusive and infrared-free methods. The second part of the article contains results of the current research on such a system. The developed human eye gaze tracking system base on simple webcam presents huge potential for further research, focused on the head motion and position analysis, precise eye gaze tracking on the screen and HCI for text input, dedicated to disabled people.**

Keywords* - *gaze tracking, gaze, IR free, eye controlled HCI

I. Introduction

Despite the active research, the eye gaze tracking remains a very challenging task due to many issues appearing during tracking process, including occlusion of the eye by the eyelids, differences in size and reflectivity, head pose, camera resolution and other technical problems. Furthermore, the most accurate solutions are expensive and unaffordable for the ordinary users.

The eye tracking brings the information not only about the gazing point of the user. Eyes reflect the way the person is thinking in a given moment, retrieving some memories or creating the new one. What is more, the eye moves, due to the complex movements programming process, can be used to detect several brain damages or diseases i.e. schizophrenia [20]. Scientific applications of eye movements tracking and analysis are vast.

Eye tracking can be also used as a human computer interface for people, who cannot operate the traditional keyboard and mouse, as well as healthy people, as an alternative way of communication with the computer [16].

II. Vision and Eye Movements Mechanisms

A. *Eye physiology*

The eye tracking applications require providing the eye model that will be used in the calculations [10]. The most typical values of the human eyes' parameters are presented in Table 1. [18].

TABLE I. HUMAN EYES' PARAMETERS

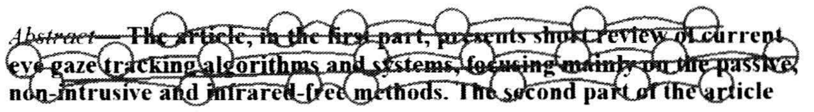
Size of eye's visual field	~135° × ~160°
Range of eyeball rotation	~70° × ~70°
Diameter of the fovea	~1 mm
Radius of the eyeball	1.3 cm

B. *Eye movements' types*

There exist three different eye movement types. The saccades are quick preprogrammed eye jumps to the next viewing point. During the "jump" the cognitive process is stopped and person becomes "blind" for a few milliseconds. In Figure 1 saccades are shown as straight lines.

The second type of eye movements are fixations, shown in the Figure 1 as circles. Fixations are small and very fast eye movements focused in one point. Presence of fixations reflects a running cognitive process and the real seeing.

The last, but not least, type of eye movements is smooth pursuits, when the eyes are smoothly and slowly following some object [19].



Abstract—The article, in the first part, presents short review of current eye gaze tracking algorithms and systems, focusing mainly on the passive, non-intrusive and infrared-free methods. The second part of the article

Figure 1. An example of saccades (curved lines) and fixations (circles)

C. *Eye movements' control*

The process of seeing is one of the most complex one. Many different parts of the brain are involved in the eye movements' control. The eye movement path is not always the same for the given stimulus. It is strongly dependent on the given task or intentions of the examined person. The cognitive process works differently when the different tasks are given. On the basis of such an eye tracking results, the cognitive processes can be understood more clearly and any defects of the whole process can be detected [19].

III. Eye Tracking Applications

Eye tracking has many different applications: from pure scientific ones, like cognitive process analysis, through medical applications, to human computer interfaces for disabled people. Due to the complexity of the eye movement programming process and involvement of many brain parts, the eye tracking can be used as a simple, fast and efficient way for diagnosis of many diseases like: schizophrenia, strokes and brain damages [19-20],

What is more, the eye tracking is used in the ergonomics and usability. Eye tracking gives an answer if the interface or the device is correctly designed and if it is intuitive. It also shows the most important parts of the interfaces, where the most significant information should be placed.

Eye tracking can be used also for fatigue monitoring and safety, for example in cars. A system can observe the road and the driver, and warn him if the unseen obstacle will appear in the eye of the cameras. In addition, the driver can be warned, when the fatigue will exceed safe level [6].

Eye tracking can be used for the remote pan-tilt-zoom camera control. The user can see the images captured by the remote webcam, while the viewing direction is affected by the eye movements [1]. Moreover, eye gaze tracking can control the avatar's eyes in the virtual environment, improving the reality of such a solutions [8].

The last described, but not the last possible, application of eye tracking is a human-computer interface for the disabled people. User could operate computer using only the eye movements [16].

IV. EYE TRACKING METHODS REVIEW

A. *Historical methods*

The first documented observations of eye movements were made in 19thcentury. The employed method relied on direct observation of the examined person's eye movements. In 1879 Louis Émile Javal has found that the reading does not require continuous eye moves with the text direction. With time, more advanced eye tracking techniques were worked out. Edmund Huey had been using a contact lens connected to an aluminum pointer. This method, despite intrusive approach to eye tracking, had proven that not all the words in the sentence are fixated [19].

Guy Thomas Buswell implemented the method that, with some improvements, is used in the modern eye tracking systems. What is even more important, it is much less intrusive than previously presented methods. Rays of the light were reflected from the surface of the eye and recorded on the video tape. Such a method assures that the tracking has hardly influenced on the experiment results. Yarbus determined that the eye movements are strongly related to the given task [19].

B. *Electrooculography*

Electrooculography is based on the measurement of the electric potential on the surface of the face near the examined person's eyes. Changes of the electric potential between the electrodes are linearly correlated with the eye movements caused by the muscles. The electric potential changes are presented in Figure 2 [19].

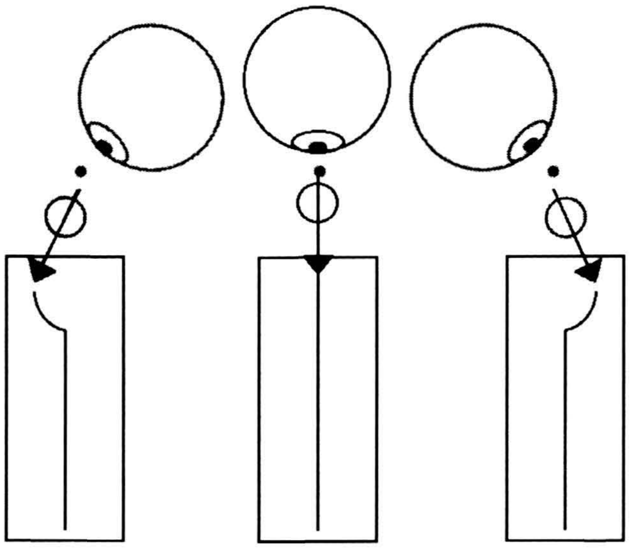


Figure 2. Electrooculography results

C. *Electromagnetic contact lens tracking*

The other eye tracking method is based on magnetic field measurements. User is wearing the contact lens with the coil that creates electromagnetic field. User's head is placed inside the frame of sensors measuring the magnetic field. The resolution of the solution is high, however, a special contact lens and complicated measurement system makes the solution impractical. Furthermore, the user can feel uncomfortable keeping his head inside the measuring frame.

D. *Photo and videooculography*

The photo/video oculography is one of the oldest eye tracking method. The camera is placed behind the glass board on which the examined person is solving the given problems. Camera can be placed also in front of the user, capturing the images of his face. The exact viewing point is determined subjectively by the operator.

E. *Reflected light method*

The reflected light method is one of the most popular eye tracking methods, due to the simple equipment, high accuracy and small influence on the examined person. It is based on the light reflections that occur on the different parts of the eyes called Purkinje reflections. Light sources are placed in the corners of the screen and near the camera lens. The light near the camera is pulsating with the changing frames. When the light near the camera is turned on the pupil is visible in the camera as a bright point, due to reflection from the retina similar to the red eyes effect. In the consecutive frame, the light is turned off, leaving the pupil dark.

The light sources placed in the corners remain turned on during the whole tracking process. The reflections that can be observed in the eye can be used for the looking direction determination and measuring the position of the reflection on the surface of the eye. First generation of the reflected light eye trackers is using the first Purkinje reflection. The more sensitive one uses both fourth and first Purkinje reflections (Figure 3) [19].

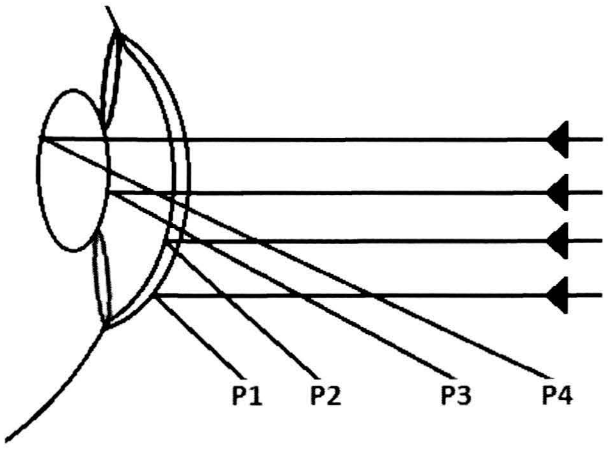


Figure 3. Purkinje reflections [http://mozyrko.files.wordpress.com]

With the change of the viewing direction, the position of the reflections on the eye will change. Reflected light eye tracking systems are very accurate and non-invasive, however, the price of the commercially available systems is very high. What is more, the continuous influence of the infrared light used in such systems can cause eye strain and be harmful for the users' eyes.

V. VISIBLE LIGHT BASED METHODS

The vision based eye gaze tracking methods can be divided into two main groups: shape-based methods and appearance-based methods. There exist also the hybrid group, employing the features of the both. The Shape-based methods use fixed and deformable shape respectively for the feature detection. The most commonly used features, in case of eye tracking, are limbus and pupil. The appearance-based approach (the holistic approach) is based mainly on the template matching. An image patch model is constructed and eye detection through model matching using a similarity measure is performed. The template matching can be performed on the intensity image or a lower dimensional subspace [4].

One of the methods for iris shape analysis is the Circular Hough Transform that is widely used in the existing visible light based methods. Kunka and Kostek employ Haar-like features classifier for the face detection and in the next step the threshold operation is performed for iris segmentation. The threshold value was found empirically during the experiments. The reference points – eye corners – are found by calculation of intensity variance changes. Irises are detected with the Circular Hough Transform [13]. Similar approach can be found in [14], where the Hough transform with gradient direction is used to perform the iris detection and gaze estimation. Another preprocessing method and reference point selection is presented in [16] and [3]. AdaBoost, Camshift and Lucas-Kanade optical flow algorithms were respectively utilized to track the face and nostrils. However, the pupils were also positioned using gradient Hough circular transform.

The other approach for the eye position and motion determination is based on initial centroid and gradient analysis technique. Moreover, the authors take into account high and low occlusion conditions, that can significantly affect the results of tracking [15]. Eye location and tracking with the use of isophotes (points with similar intensity value) properties is reported in [7]. The presented method is robust to linear lighting changes and rotational invariance. Li and Parkhust adapted the Starburst algorithm, finding the limbus contour points by computing the derivatives along rays extending radially from the initial point, until a threshold derivative value is obtained [11].

The "one-circle" algorithm for measuring the eye gaze uses an image of only one eye. Observing that the iris contour is a circle, authors estimate the normal direction of this iris circle, considered as the eye gaze, from its elliptical image. Such an approach brings two solutions for the projections, however the geometric constraint (the distance between the eyeball's center and the two eye corners should be equal to each other) removes one of the solutions [9].

Eye gaze tracking can be also based on the Active Appearance Model (AAM). Authors use AAM for the whole head tracking. Required features (eye corners, eye region) are extracted from the whole head model [6].

One of the main constraint of the eye gaze tracking using simple webcam is the camera resolution. Strictly speaking, the dimensions of eye image extracted from the whole frame. One of the possible solutions of the problem is subpixel tracking method presented in [5]. Eye corners and irises contours are interpolated with 1-dimensional cubic interpolation, giving much higher accuracy of the tracking.

Eye tracking requires a reference points selection that usually are eye corners or nostrils. Instead of these, a small 2D mark place on the user's face can be employed to compensate for the head movements [12].

VI. EXISTING METHODS SUMMARY

The presented techniques have their advantages and limitations, but the optimal performance of any technique also implies that its optimal working conditions are met. These conditions relate to illumination, head pose, ethnicity, and degree of eye occlusion. For example, the outdoor illumination affects Infrared based tracking methods, while techniques based on shape and appearances can work both indoors and outdoors [4].

The existing methods, besides the visible light based methods, poses many disadvantages that can influence the examination process. The main disadvantage of almost all presented solutions is high price of the system. As the eye tracker is used as laboratory equipment, the price is not a crucial factor. However, such systems can be used as human computer interfaces for the disabled people, for which low price is an important feature of the equipment used.

The other disadvantage is invasiveness of the method. Some of the methods require usage of special contact lenses, while others uses artificial light sources directed straight into the users eyes. The influence of contact lenses or the light can change the results of the examination, disturbing the examined person or causing uncomfortable sensations. Moreover, the continuous usage of the artificial light sources directed into the user's eyes can be harmful for the user.

VII. Proposed solution

A. Previous research

The first version of the developed eye gaze tracking algorithm was based on Haar-like features classifier for the face and eyes recognition and basic threshold operation for pupils' detection and reference point calculation. Due to the specific operation of the Haar-like features classifier, the middle point of the face is not stable and cannot be used as a reference point for viewing direction determination.

To avoid fluctuations of the middle point of the face calculated from the classifier's output, the two stage threshold operation was implemented in the application. The purpose for the first operation was to detect pupils of the eyes, as darkest areas in the image. While there often exist many other dark points in the image (i.e. eye corners), only the two largest regions were taken into account. The selection of these regions is based on contours area calculation. Contours are segmented from the binary image with the use of the border following algorithm proposed by Satoshi Suzuki [17]. The second threshold operation was used to remove points that were significantly brighter than the eye pupils, leaving the approximate area of the eye sockets.

The next step was the center point calculation for each obtained contour – two for eyes and one for the whole eye sockets area. Assuming that, the center point of eye sockets is a reference point, all eye pupils' movements were determined with respect to this point.

B. Current solution

1) Overall algorithm review

The main purpose of the application is the human eye gaze tracking. While the problem of eye tracking is a complex one, the modular design of the application was used. It allows modules to be exchanged during the process of development. All the parts of the system are described in details further in the article.

2) Image enhancement methods

Image obtained by the web camera requires often some preprocessing stages to assure a proper quality for further analysis. In the described system, acquired images are converted from the RGB colour components to gray scale and then the image histogram is calculated and normalized. The algorithm normalizes brightness and increases image contrast.

3) Face detection and region of interest extraction

The next step of the processing is the face detection in the image. There exist many different approaches for the face detection. Some are based on the symmetry of the face and geometric relations between face features (knowledge-based). The information that can help while face detection can be skin color (feature-based) or the correlation between the training set of face patterns and the examined face (template matching). Other face detection algorithms employ also neural networks.

In the presented application the Haar-like features classifier is used, while it is one of the most accurate, robust and fast method for detecting features. The principle of operation of the classifier is based on simple features calculations and cascading many simple and weak classifiers to obtain a strong one.

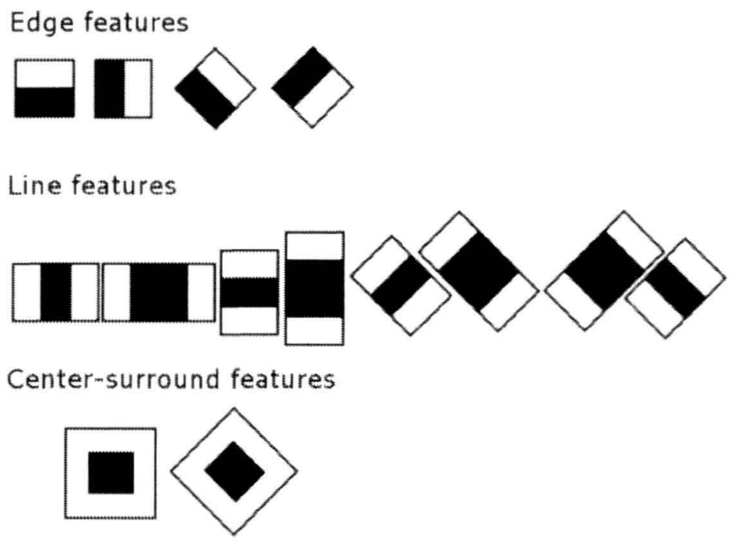


Figure 4. Haar-like features

Features, presented in Figure 4, cover parts of the image. The value of the feature in the given image location is calculated as a weighted difference of pixels covered by the whole feature and pixels covered by the black area. Each feature mask is scaled and moved through the whole image during face detection. Obtained values are an input for a tree of classifiers, determining if the part is face or not. The adaptive boosting process (AdaBoost) is implemented, to increase the recognition rate. AdaBoost finds the classifiers that minimizes the error rate and updates (increases) weights of incorrectly classified points that are considered as more important in the next iteration. The process repeats until error rate is higher than 0.5. The boosting process is repeated several times to build a cascade of classifiers [21].

Each classifier after operation of adaptive boosting has high true positive detection rate (about 0.999), but also very high false positive rate (about 0.5). Using only one of such classifiers will make no sense; however, composing them into the cascade creates efficient classifier, as shown in equation

$$0.999^{20} \approx 98\% \; \textit{for true positive}, 0.5^{20} \approx 0.0001\% \; \textit{for false positive} \quad (1)$$

4) Face's center point calculation and tracking

The center point of the face, calculated using the data from the Haar-like features classifier, is fluctuating, due to the specificity of the algorithm. During computation, each feature is scaled to fit the input image. If the scaling step is large, the face detection is fast but inaccurate. In consecutive frames face can "jump" from one point to other, even if the user is not moving his head. In the contrary, the scaling factor close to 1 decreases substantially the speed of processing.

In order to obtain a real-time computation speed as well as stable detection of the center point of the face, the middle point between eye corners is detected as it is shown in Figure 5. Starting from the middle vertical line of the face that is located between the eyes, the two darkest points placed on the left and right side are found. The program assumes that these points are the eye corners. Calculating the average from points' position, the center point of the face is obtained. Stability of such an obtained point is sufficient for the eye tracking purposes.

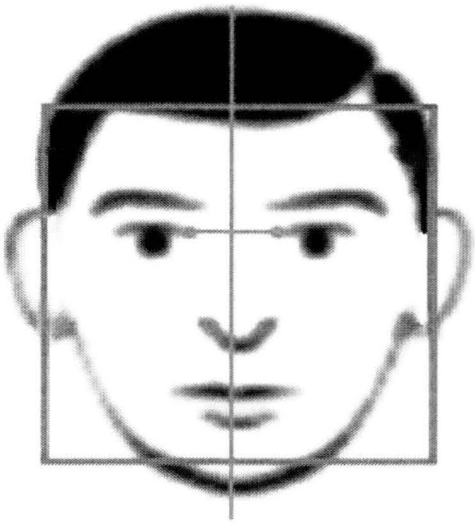


Figure 5. Eye corners detection

5) Pupil detection

Recognition of the eyes on the image is based also on the Haar-like features classifier. The only difference is that it was trained on the eye images data base. Having the rough regions for each eye, the exact pupil shape should be calculated. In each region the largest set of the dark pixels are assumed to be a pupil region. For users with dark eye colors, such a region will be a pupil with the iris. The center point of each pupil is calculated as a center of a gravity of each region contour. Such an approach was sufficient in the early phase of the project development, but it will be improved to obtain better resolution and accuracy.

6) Viewing point calculation

Due to the instability of the user's face position, the pupils' positions are calculated with reference to the center point of the face as was described in point 4. Obtained vectors represent the relative displacements of each eyeball.

7) Tracking results

The eye gaze tracking results obtained with the designed system are presented in Figures 6 and 7 The lines represents the eye movements with reference to the center point of the head.

During the experiment, the user was asked to make 4 eye movements: up, down, left and right. The distance between head and screen with the camera was equal about 40 cm. No additional light sources were used.

The vertical eye movements tracking results are presented in Figure 6. Two stars represent the eye gaze points in the screen and the corresponding tracking result (arrows).

The horizontal eye movements can be seen in Figure 7. Straight position of the eye is marked with number 4, while the marginal positions with 3 and 5. Point no 6 is the result of the tracking failure, however such errors can be detected and taken into account during the final processing of the eye gaze direction.

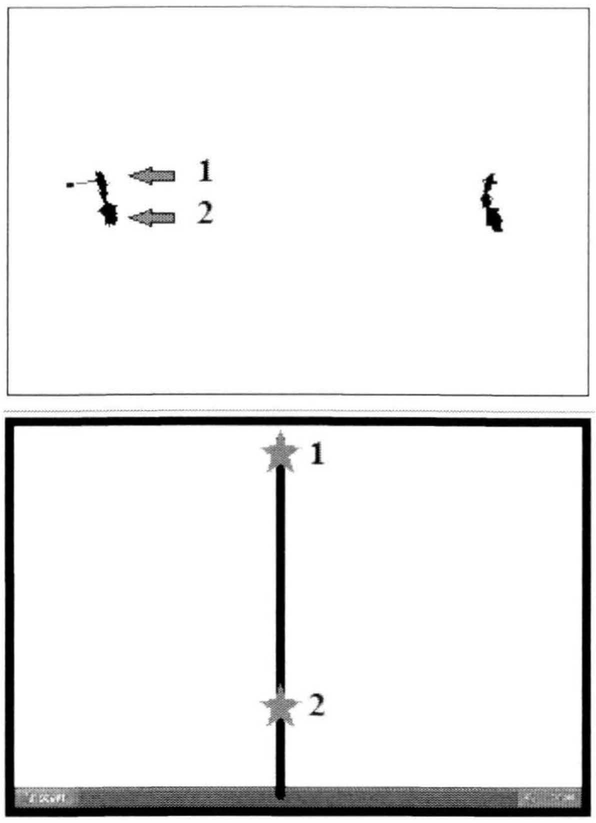


Figure 6. Tracking results of vertical eye gaze movements

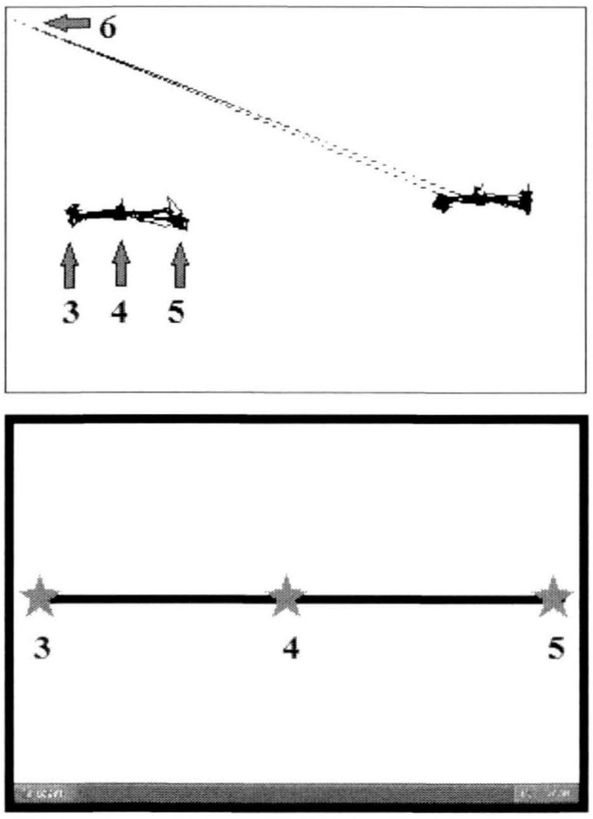


Figure 7. Tracking results of horizontal eye gaze movements

As it is visible in Figures 6 and 7, system shows acceptable accuracy and precision, however, the vertical movements are less noticeable, while the vertical eye movements have narrower range then horizontal ones.

VIII. Conclusions

Current version of the system is able to show the relative eye movement of the user. The accuracy of operation and system robustness require further improvements, however it is possible to build the passive real-time eye tracking system using simple webcam, possessing relatively high precision and resolution. Such a system will be a perfect solution in situations when additional light source (particularly Infrared LEDs) can be dangerous or harmful, i.e. for outdoor applications and HCI systems for persons with disabilities, where the ergonomics and price of the system plays a key role.

IX. REFERENCES

[1] J. Chen, J. Han, T. Lan and X. Li, "The real-time eye-controlled system based on the near-field video streams," in International Symposium on IT in Medicine and Education, Cuangzhou, 2011.

[2] Z. Ramdane-Cherif and A. Nait-Ali, "An adaptive algorithm for eye-gaze-tracking-device calibration," IEEE Transactions on Instrumentation and Measurement, vol. 57, no. 4, pp. 716 - 723, 2008.

[3] B. L. Nguyen, "Eye gaze tracking," in International Conference on Computing and Communication Technologies, Paris, 2009.

[4] D. W. Hansen and Q. Ji, "In the eye of the beholder: a survey of models for eyes and gaze," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 2, pp. 478 - 500, March 2010.

[5] J. Zhu and Y. Jie, "Subpixel eye gaze tracking," in Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Pittsburgh, 2002.

[6] T. Ishikawa, S. Baker, I. Matthews and T. Kanade, "Passive driver gaze tracking with active appearance models," in Proceedings of the 11th World Congress on Intelligent Transportation Systems, 2004.

[7] R. Valenti and T. Gevers, "Accurate eye center location and tracking using isophote curvature," in IEEE Conference on Computer Vision and Pattern Recognition, Amsterdam, 2008.

[8] N. Murray, D. Roberts, A. Steed, P. Sharkey, P. Dickerson, J. Rae and R. Wolff, "Eye gaze in virtual environments: evaluating the need and initial work on implementation," Concurrency and Computation: Practice and Experience, no. 21, p. 1437–1449, 2009.

[9] W. Jian-Gang and E. Sung, "Study on eye gaze estimation," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 3, no. 32, pp. 332 - 350, 2002.

[10] M. Che, B. Zhang, K. Cen and W. Gao, "A novel simple 2D model of eye gaze estimation," in International Conference on Intelligent Human-Machine Systems and Cybernetics, Tianjin, 2010.

[11] D. Li and D. Parkhurst, "Open-source software for real-time visible-spectrum eye tracking," in The 2nd Conference on Communication by Gaze Interaction, Turin, 2006.

[12] K. Kyung-Nam and R. Ramakrishna, "Vision-based eye-gaze tracking for human computer insterface," in International Conference on Systems, Man, and Cybernetics, Tokyo, 1999.

[13] B. Kunka and B. Kostek, "Non-intrusive infrared-free eye tracking method," in Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, Poznań, 2009.

[14] X. Sun, G. Chen, C. Zhao and J. Ya, "Gaze Estimation of Human Eye," in International Conference on ITS Telecommunications Proceedings, Chengdu, 2006.

[15] S. Wibirama, S. Tungjitkusolmun, C. Pintavirooj and K. Hamamoto, "Real time eye tracking using initial centroid and gradient analysis technique," in International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Pattaya, 2009.

[16] W. Tunhua, B. Baogang, Z. Changle, L. Shaozi and L. Kunhui, "Real-time non-intrusive eye tracking for human-computer interaction," in International Conference on Computer Science and Education, Hefei, 2010.

[17] Suzuki, S. and Abe, K., "Topological Structural Analysis of Digitized Binary Images by Border Following". CVGIP 30 1, pp 32-46, 1985

[18] B.A. Wandell, "Foundation of Vision." Sinauer Press, Sunderland, MA, 1995.

[19] A. Duchnowski, "Eye tracking methodology: theory and practice." Springer-Verlang, 2003.

[20] Ross, David E;Buchanan, Robert W;Medoff, Deborah;Lahti, Adrienne C;Thaker, Gunvant K "Association between eye tracking disorder in schizophrenia and poor sensory integration," The American Journal of Psychiatry; Oct 1998; 155, 10; ProQuest pp 1352-1357

[21] Viola P., Jones, M.: Rapid object detection using a boosted cascade of simple features, Computer Vision and Pattern Recognition, vol. 1, pp. 511 518, 2001

Use of Haar-like Features in Vision-Based Human-Computer Interaction Systems

Aleksandra Królak
Institute of Electronics
Technical University of Łódź
Łódź, Poland
aleksandra.krolak@p.lodz.pl

Abstract — **In everyday life human gestures are often used to communicate or enhance speech. They can be also used to enable or improve the communication between human and machine. Among the contactless human-computer interfaces (HCI) the vision-based solutions enabling for face and hand gesture recognition are the most promising ones. Haar-like object detection algorithm developed by Viola and Jones allows for rapid detection of human faces or hands in image sequences. This paper presents the overview of vision-based Human-Computer interfaces employing methods based on Haar-like features and proposes a Human-Computer Interaction system controlled by mouth shape change.**

Keywords — Human-Computer Interaction (HCI), Haar-like features, face detection

I. Introduction

Over the years Human Computer Interfaces (HCI) evolved from text through graphical to multimedia ones. However the most common input method is still by using computer keyboard and mouse. Unfortunately these devices are not sufficient to meet the needs of the latest virtual reality applications. They are also not suitable for motorically impaired users. Therefore the development of alternative methods of communication between human and computer attracts the interest of many researchers in recent years. Two trends can be observed in this field: research on possibly most natural ways of human-computer interaction, and building systems enabling communication with the computer for people with severe physical disabilities. In both cases two types of solutions are utilized: systems with external devices mounted on user's body, and contactless systems which offer much more comfort for the users. The non-intrusive systems utilize different types of remote sensors but the most promising ones are vision-based solutions. User friendly human-computer interface should fulfill several conditions: be contactless, not dependent on lighting conditions, be reliable and working in the real time.

Haar-like object detection algorithm developed by Viola and Jones [1] allows for rapid and reliable detection of human faces or hands in image sequences. Since the classifiers build using this approach are available as open source, this method is probably the most frequently used one in the development of vision-based Human-Computer Interfaces.

In this paper the Haar-like object detection method is briefly explained, followed by the overview of Human-Computer Interaction systems based on Haar-like classifier for face features detection. Finally a mouth shape controlled vision-based HCI is proposed and the results and conclusions are presented.

II. Haar-like object detection

Face detection in image sequences is a challenging problem and many techniques have been developed to solve it. The existing approaches can be classified into four groups:

- knowledge-based methods employing simple rules to describe the properties of the face symmetry and the geometrical relationships between face features [2];
- feature-based methods based on the detection of mouth, eyes, nose or skin color [3, 4];
- template matching methods based on finding the correlation between the input image and stored patterns of the face [5];
- appearance-based methods, where algorithms are trained on models using neural networks [6], Support Vector Machines (SVM) [7] or Hidden Markov Models (HMM) [8].

Haar-like object detection is a method derived from the template matching group. It was developed by Viola and Jones [1] and modified by Leinehart and Maydt [9]. The algorithm allows for rapid detection of any type of the object in the image for which the classifier is trained. The AdaBoost classifier cascades are based not on pixel intensities but on so called Haar-like features.

Haar-like features enable encoding different features of an image by encoding contrasts exhibited by the object of interest and their spatial relationship. Each Haar-like feature can be considered as a template of several white and black rectangles interconnected. Haar-like features are computed similarly to the coefficients of transformations based on Haar wavelet. The features used are rectangular and of different size, subdivided into white and black regions. The three types of Haar-like features are presented in the fig. 1.

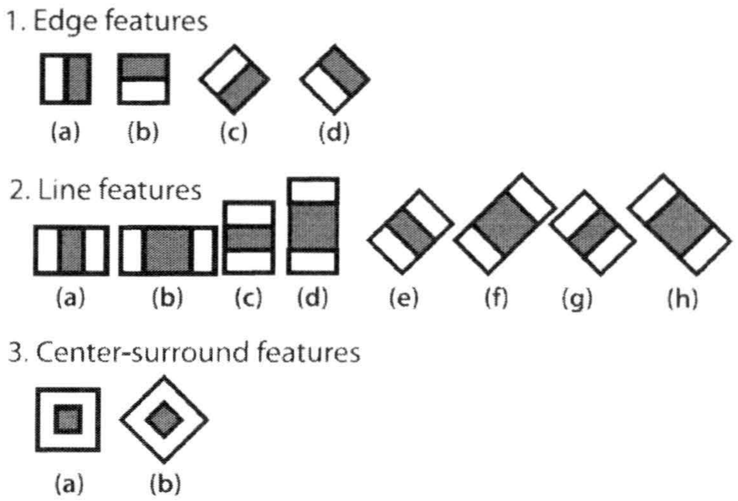


Figure 1. Rectangular masks used for object detection

The feature value for the given mask is calculated as the weighted sum of the intensity of the pixels covered by black rectangle and the sum of pixel intensities covered by the whole mask. Defining s_1 as the black region of the mask and s_2 as the whole mask, A_1 as the area of the black rectangle of the mask and A_2 as the area of whole mask, the weights of these two areas respectively can be defined as:

$$w_1 = -1. \quad (1)$$

$$w_2 = A_1/A_2. \quad (2)$$

The value c of the feature is calculated according to the equation:

$$c = w_1 s_1 + w_2 s_2. \quad (3)$$

where s_1 and s_2 are the sums of pixel intensities covered by the black rectangle and the whole mask respectively.

The key advantage of a Haar-like feature over most other features is its calculation speed. Due to the use of integral images, a Haar-like feature of any size can be calculated in constant time (approximately 60 microprocessor instructions for a 2-rectangle feature) [1].

Integral image is defined as a 2-dimensional lookup table in the form of a matrix of the same size as the size of the original image. The value of the matrix at position (x,y) is the sum of pixel intensities above and to the left of (x,y) inclusive. This allows for computing the sum of rectangular areas in the image, at any position or scale, using 4 lookup tables only. For the extended set of Haar-like masks the features are computed using 6 lookup tables for 2-rectangle masks, 8 lookup tables for 3-rectangle masks and 9 lookup tables for 4-rectangle masks. The sum of pixel intensities for the rectangle defined by points *pt1*, *pt2*, *pt3* and *pt4* is calculated form equation (4), here A, B, C and D are areas as shown in fig. 2, and E is the sum of regions A, B, C and D.

$$D = E + A - B - C. \quad (4)$$

Single features are used by the large set of the weak classifiers to label the particular image region as "object" or "non-object".

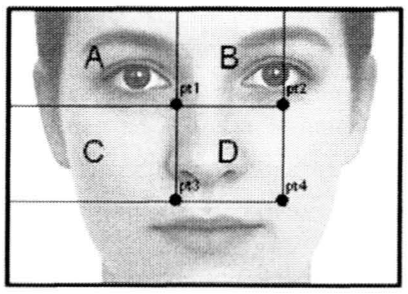


Figure 2. Integral image calculation

Not all calculated features are necessary to correctly detect desired objects in the image. Effective classifier may be formed using only a part of the features with smallest error rates. In order to find these features the boosting algorithm AdaBoost is used. As a result a cascade of boosted classifiers is built. For face detection and face feature detection the obtained accuracy rates are presented in Table 1 [15]. It can be stated that Haar-like feature based methods have very high rate of detection (over 90%), but the false positive rate is also relatively high (about 25%).

TABLE I. FACE FEATURE DETECTION ACCURACY

Object	Accuracy
Face	95%
Eyes	80%
Nose	78%
Mouth	71%

III. HAAR-LIKE FEATURES IN VISION-BASED HCI SYSTEMS

Haar-like object detection is a method widely used in the vision-based human-computer interaction systems. A number of vision-based interaction systems were developed using this approach, that enable the communication with the computer by performing hand gestures [10], eye blinking [11], mouth movements [12] or head movements [13].

Quite simple but very useful solution was developed by researchers from Chongquing University [13]. They proposed a vision-based system for controlling the intelligent wheelchair by head movements. For this purpose a lip detector based on Haar-like features was used. The assumption made is that the head of the user is located in the middle of the image from the webcam. The head position changes are determined by the relative changes of the mouth location with respect to the center of the image. The system is able to recognize four head positions (apart from the neutral one): head up, head down, turn left and turn right. These changes of head position trigger the wheelchair actions: go forward, go backward, turn left and turn right respectively. In the paper no numbers concerning the accuracy of the system are reported, however it is expected at the level of about 70% as the average accuracy of the cascade classifier for mouth detection.

Another system where the mouth position is analyzed is a lip movement multimodal human-computer interface proposed by a research group from Gdansk University of Technology [12]. The system is designed especially for severely disabled and paralyzed users to enable them control over computer mouse and keyboard. The position of mouth is determined not with the cascade classifier trained for mouth detection but the Haar-like face detector. The mouth region s localized in the lower part of the detected face region. Further lip shape estimation is performed in the image composed of components of LUV color space and Discrete Hartley Transform (DHT). Recognition and classification of the mouth shape is done using neural network. The system allows for recognition of four mouth shapes: neutral, opened, "O" shape and sticking out the tongue. The average effectiveness of the LipMouse is about 85%.

Human-computer interface controlled by different type of face gestures, that is by eye blinks, is b-Link invented at the Lodz University of Technology [11] and brought to market by the Orange group. In this bimodal interface two Haar cascades are used: for face detection and for eye detection. Face detector is used to minimize the region of interest for further eye detection. The blinks are detected by template matching, that is the current eye image is compared with the template eye image of the user acquired at the initialization of the system. In order to distinguish the voluntary (control) blinks from the spontaneous ones only eye-blinks lasting from 250ms to 1s are considered as the input signals to the system. The accuracy of this eye-blink controlled human-computer interface is reported to be ~95%. The system, designed especially for motorically impaired persons, offers the user many functions, such as on-screen keyboard and mouse, menus with shortcuts to favorite websites and applications and possibility of turning off the computer.

Solutions for the disabled include not only HCI systems but also emotion recognition systems. EmoCam [14], developed in the Swinburne University of Technology, is designed for persons with physical disabilities and persons having problems with expressing feeling verbally (cases of autism, cerebral palsy or speech impairments). The system composed of a laptop computer and a webcam allows for recognition in the real time of five basic emotional states: neutral, angry, happy, sad and scared. The first stage of the proposed algorithm is face detection using Haar-like features. It is followed by Principal Component Analysis that uses Eigenfaces to detect emotions.

Alternative methods of human-computer interaction are designed not only to enable the disabled users communication with the machine but also to allow more natural interaction in the virtual reality environments. An example of such solution is hand gesture recognition system developed at the University of Ottawa. This system enables detection of four hand poses: "two fingers", "palm", "fist" and "little finger". The hand gestures are recognized in two stages: posture recognition using parallel architecture of four cascade classifiers based on Haar-like features, and gestures recognition using the syntactic analysis based on stochastic context-free grammar. The accuracy of the real time detection of these hand postures is reported to be about 95%.

IV. PROPOSED BIMODAL HCI

The proposed mouth-controlled HCI is based on the graphical user interface designed for b-Link system [11]. The change of the input method was suggested by one of the disabled users of eye-blink controlled interface who is suffering from athetosis. For persons with this type of disability voluntary eye-blinks are quite difficult to perform, while the gesture of showing teeth does not cause problems. Therefore the proposed HCI offers the users the same functionality as b-Link.

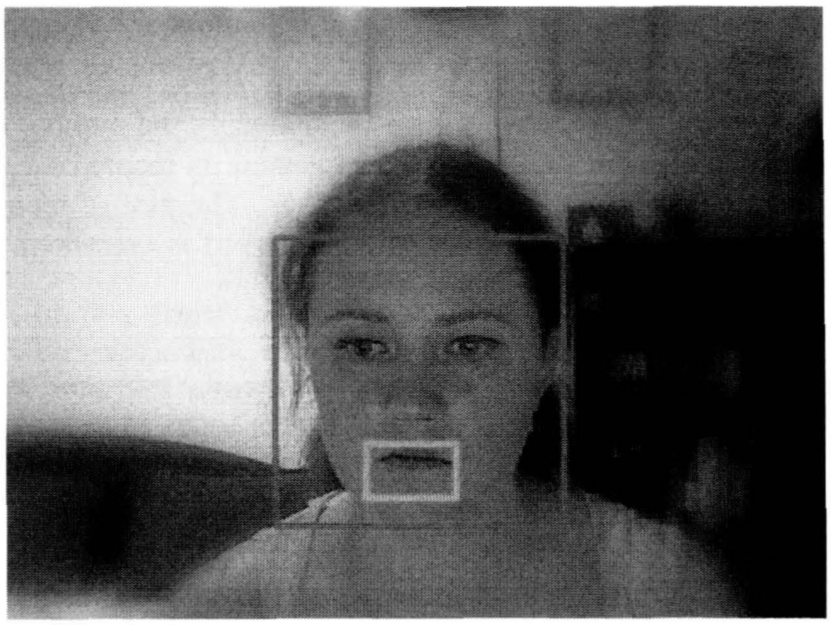


Figure 3. Result of face and mouth detection

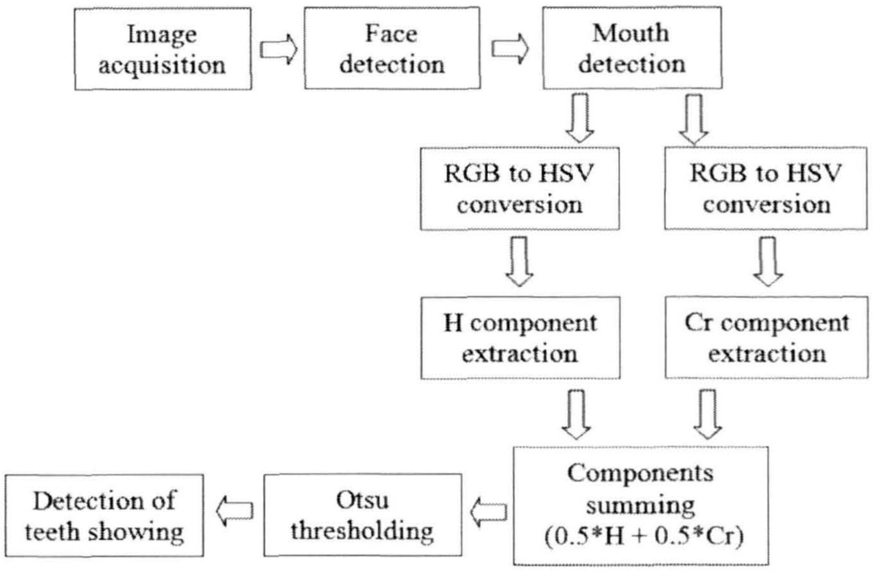


Figure 4. Block dagram of the proposed HCI system

The first step of the algorithm is Haar-like face detection which allows for reducing the size of image region for mouth detection. The mouth region detected is also performed with the cascade classifier. The resulting image with detected face and mouth is shown in fig. 3. The output image ROI is analyzed in two color spaces: HSV and YCbCr. The hue component (H) and red chrominance (Cr) component are added with equal weights and the resulting image is thresholded using Otsu method [16]. The block diagram of the system is presented in fig. 4. The subsequent stages of image processing for closed mouth and the control mouth shape are shown in fig. 5.

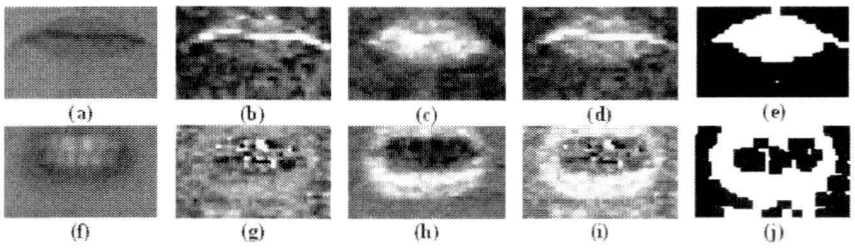


Figure 5. Stages of mouth shape detection: RGB image (a, f), hue component (b, g), red chrominance component (c, h), summed image (d, i), thresholded image (e, j).

The detection of the desired mouth shape, that is the gesture of showing teeth, is done by analyzing the number of white pixels in the final image. The gesture is recognized as the control one if it lasts for at least 0.5s. The plot of mouth shape are in time with trigger actions marked is presented in fig. 6.

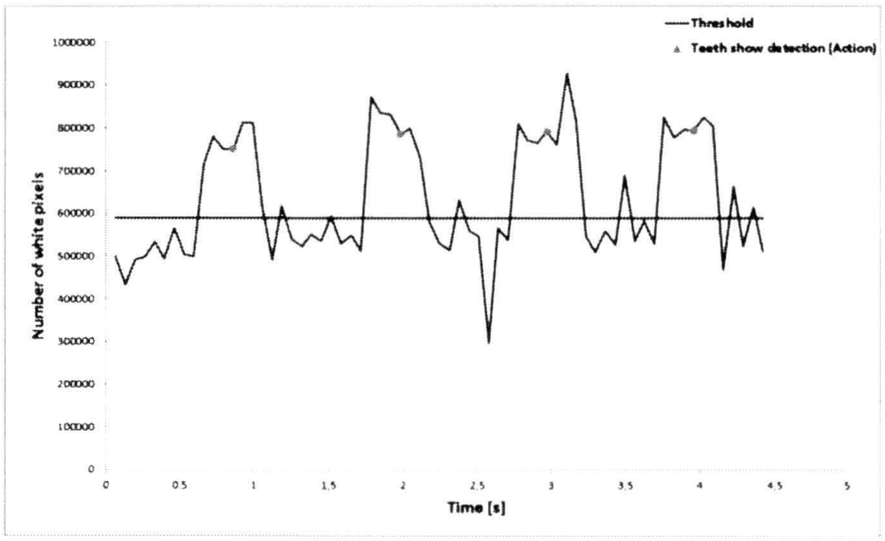


Figure 6. Plot of the number of white pixels vs. time with control gesture detection

The system was tested by 12 users, 3 female and 9 male. Two of the male users had facial hair. Each user was asked to make 10 control mouth gestures. The rate of correct recognition of the desired mouth gesture was equal to 96%. The tests included also monitoring of the system behavior during speech conversation. The results showed that mouth gestures made during regular speech do not influence the performance of the proposed system.

V. CONCLUSIONS

Presented overview shows that Haar-like object detection is often successfully used in vision-based interaction systems. The reason for it is the reliability, high accuracy and speed of this method. A great advantage of this approach is the fact that it is not susceptible to noise and light conditions, and is scale invariant. This method is often used in the object detection and recognition systems also due to the fact that most of the trained cascade classifiers are available as open source.

Proposed Human-Computer Interface controlled by "showing teeth" gesture is characterized by high detection accuracy (~96%). This result suggests that such input method can be successfully employed in the bimodal HCI. It can be used as an extension for the eye-blink controlled vision-based human-computer interface [11].

REFERENCES

[1] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society, vol. 1, pp. 511- 518, 2001

[2] G. Yang, T. S. Huang, "Human face detection in complex background", Pattern Recognition, vol 27(1), pp.53-63, 1994

[3] K. C. Yow, R. Cipolla, "Feature-based human face detection", Image and vision computing, vol 15(9), pp.713-735, 1997

[4] S. McKenna, S. Gong, Y. Raja, "Modelling facial colour and identity with gaussian mixtures", Pattern Recognition, vol 31(12), pp.1883-1892, 1998

[5] C. Lanitis, J. Taylor, T. F. Cootes, "An automatic face identification system using flexible appearance models", Image and Vision Computing, vol. 13 (5), pp. 393-401, 1995

[6] H. A. Rowley, S. Baluja, T. Kanade, "Neural Network-Based Face Detection", IEEE Transactions On Pattern Analysis and Machine Intelligence 20(1), pp. 23-38. 1998

[7] E. Osuna, R. Freund, F. Girosi, "Training Support Vector Machines: an application to Face Detection", Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.130-136, 1997

[8] K. Rajagopalan, K. Kumar, J. Karlekar, R. Manivasakan, M. Patil, U. Desai, P. Poonacha, S. Chaudhuri, "Finding faces in photographs", Proc. sixth IEEE International Conference on Computer Vision, pp.640-645, 1998

[9] R. Leinhart, J. Maydt, "An extended set of Haar-like features", Proc. Int. Conf. on Image Processing, pp. 900-903, 2002

[10] Q. Chen, N. Georganas, "Hand Gesture Recogniton Usng Haar-Like Features and a Stochastic Context-Free Grammar", IEEE Transactions on Instr. and Measuremen, vol. 57 (8), pp. 1562-1571, 2008

[11] A. Krolak, P. Strumillo, "Eye-blink detection system for human-computer interaction", International Journal on Universal Access in the Information Society, vol. 10, 2011, DOI 10.1007/s10209-011-0256-6

[12] P. Dalka, A. Czyzewski, "Lip movement and gesture recognition for multimodal human-computer interface", Proc. of the International Multiconference on Computer Science and Information Technology, pp. 451-455, 2009

[13] Z.-F. Hu, L. Li, Y. Luo, Y. Zhang, X. Wei, "A Novel Intelligent Wheelchair Control Approach Based On Head Gesture Recognition", Proc. of ICCASM, pp159-163, 2010

[14] B. T. Lau, "Portable real time emotion detection system for the disabled", Expert Systems with Applications, vol. 37, pp 6561-6566, 2010.

[15] P. I. Wilson, J. Fernandez, "Facial Feature Detection Using Haar Classifiers", Journal of Computing Sciences in Colleges archive, vol. 21(4), pp. 127-133, 2006

[16] N. Otsu, "A threshold selection method from gray-level histograms", IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, 1979.

A Vision-Based Head Movement Tracking System for Human-Computer Interfacing

Paweł Strumiłło
Institute of Electronics, Technical University of Lodz
211/215 Wólczańska, 90-924 Lodz, Poland
e-mail: pawel.strumillo@p.lodz.pl

Tomasz Pajor
Institute of Electronics, Technical University of Lodz
211/215 Wólczańska, 90-924 Lodz, Poland
e-mail: tomasz.pajor@piproject.pl

Abstract — **In this paper we report a vision-based human-computer interface that enables touch-free communication with a computer by means of head movements and eye-blinks. The developed interface is intended to be used by physically disabled persons. It was shown that image analysis software running on commercial off-the-shelf hardware (a mid-range notebook and a web-camera) allows for successful implementation of typical computer interaction tasks, such as: browsing the Internet, viewing image albums and reading pdf files. Test results of the interface with participation of 6 volunteers are summarized. The paper also includes a short review of other up-to-date solutions to human-computer interaction for the disabled, ranging from brain-computer interface to a tongue control system.**

*Keywords — **human-computer interaction, soft computing, the disabled, image processing and analysis***

I. Introduction

A feeling of exclusion from society and the lack of independence are a common problem for persons with physical disabilities [1]. This is particularly the case of persons suffering from partial or total paralysis (e.g. due to stroke, injury or illnesses). The physically disabled can be fully fit mentally, yet unable to communicate with the environment, a condition known as the locked-in syndrome. Recent findings of fMRI studies (functional Magnetic Resonance Imaging) indicate that such persons, although unable to perform any body control action, frequently regain awareness of self. Patients with the locked-in syndrome can perceive the environment by sight, hearing or touch. It is also argued that they understand speech and form knowledgeable mental responses, that are reflected by activation of the corresponding brain regions [1]. A large number of such persons also recover fractional control over their body, e.g. they can move their eyes and blink or perform limited head movements.

Recent advances in ICT (Information and Communication Technologies) and artificial intelligence allow to build devices that offer novel and inexpensive solutions for interacting with computers. Such devices fall into the category of human-computer interaction (HCI) technologies. A number of worldwide and European initiatives have been launched to advance work on HCIs, e.g. the COGAIN network of excellence - a European project that integrates efforts of the research community on gaze interaction systems (www.cogain.org). A number of conferences are devoted to the area of HCI, e.g. the International Conference on Computers Helping People with Special Needs will take place for the 13th time. A new journal - the ACM Transactions on Interactive Intelligent Systems, was also recently released (http://tiis.acm.org).

In this paper a HCI solution is reported that requires commercial off-the-shelf (COTS) computer components only. The interface allows for touch-free communication with the computer by sequences of head movements and eye blinks.

A. Short overview of human-computer interfaces for the disabled

Standard computer input devices such as the keyboard or the mouse are inaccessible for people with severe mobility impairments. Custom solutions are necessary to enable access to computers for the physically disabled. The approaches can be subdivided into the three following groups:

- vision-based systems, e.g. eye-blink monitoring systems,
- systems based on non-conventional mechanical interfaces or sensors, e.g. mouth operated joysticks
- systems based on electrical measurements and analysis of biosignals, e.g. brain computer interfaces (BCI).

Within the group of vision-based interfaces are e.g.: EyeTech – a mouse tracking device [2], B-link [3] and I4Control® [4] which are systems for monitoring eye-blinks. The main advantage of these systems is that they are contact-less.

The EyeTech system is a gaze tracking mouse replacement kit. It consists of a pair of infrared (IR) light sources and a camera. The camera lens is focused on the user's face. EyeTech comes with a software package that runs on-line, detects the pupil positions and calculates the corresponding gaze point on the screen. Clicking is selected from a menu (left, right, double click, etc.) and activated with either blinking or staring at a fixed position for a preset amount of time.

The B-link is a vision-based system that utilizes COTS (commercial off-the-shelf) hardware. It monitors a user's eye blinks and recognizes the voluntary ones (i.e. lasting longer than 100 ms), interpreting them as control commands. The primary use of B-Link is to operate an on-screen keyboard, which highlights keys in rows and columns to be accepted by eye blinks. Web browsing functionality is also provided. The application designed at the Lodz University of Technology was deployed in cooperation with Orange Labs and is distributed free of charge as open source software (http://sourceforge.net/). Trained users achieve the typing rate of 10 characters per minute.

Another vision based human-computer interface is the Cyber-eye system developed at the Gdańsk University of Technology. The system employs infrared light to facilitate the camera to track eye-gaze of a user who locates keys on a virtual keyboard displayed on a computer screen.. The system offers also EEG measurement. Results of EEG analysis are shown to the user by face gestures of an avatar displayed on a computer screen (sound.eti.pg.gda.pl/news/media).

The I4Control® system was developed at the Czech Technical University of Prague. The system monitors a user's eye by a tiny camera attached to user's spectacle frames, i.e. gaze tracking is achieved by the videooculography method. It was demonstrated that the system enables the physically handicapped to enter text (via a virtual keyboard) and browse the internet. The disadvantage of the systems is that an extra interface hardware connecting the camera to the computer is required.

Good examples of non-conventional mechanical devices and sensors are: a head pointer [5], Jouse – a mouth operated joystick [6] and Tongue Control (TC) [7].

The head pointer device consists of a rigid rod on a harness strapped to a user's head. The rod is protruding forward so that the tip is in the user's field of view and can be used to press a touch screen or a keyboard. Jouse is a mouth operated joystick that replaces the ordinary hand operated mouse. Cursor position is altered by moving the joystick with either mouth, tongue, chin or cheek. Mouse clicks are performed by inhaling or exhaling through a tubular mouthpiece. Use of Jouse and other joystick based systems does not require attaching any accessories to the user's body.

The TC device is a tongue-computer interface which consists of 18 inductive sensors placed within the user's mouth on a small hard palate and a ferromagnetic bead glued to the tongue. The tested individuals achieved typing rates of up to 70 characters per minute after 3h training sessions.

Finally, among the systems that are based on electrical measurements and analysis of biosignals the example solutions are: Nessi [8] – an EEG-Controlled Web Browser for Severely Paralyzed Patients and a BCI interface [9] that detects brain electrical responses to periodic light flashes.

Nessi is a Mozilla based web browser that uses signals gathered by the so called thought translation device (TTD) developed at the University of Tubingen, Germany. TTD measures slow cortical potentials (SCPs), i.e. signals a person can learn to control and invoke voluntarily. The browser displays a visual feedback containing two square fields, colored red and green, that correspond to one of two possible actions selectable in the remaining part of the screen.

The BCI interface built at the Technical University of Lodz operates on the principle of detecting Steady State Visually Evoked Potentials (SSVEP). The user observes a keyboard of flashing LEDs, with unique frequencies, ranging from 3,5 Hz to 75 Hz, corresponding to each character. The recorded EEG potentials feature those characteristic frequencies, thus it is possible to determine the character which has the user's focus. An average typing rate of 5 sec per character is currently being achieved by the individuals who tested this system.

Functional properties and availability of the reported human-computer interfaces are compared in Table 1.

TABLE I. COMPARISON OF THE SELECTED HUMAN-COMPUTER INTERFACES WITHIN THE REPORTED CLASSES

Assistive Device	Assistance to initialize	Elements attached to body	Oral contact	Zero-force operation	Difficulty of use	Price
EyeTech TM3*	YES	NO	NO	YES	LOW	10kUSD
B-link*	YES	NO	NO	YES	LOW	open source
Cyber-eye*/***	YES	NO	NO	YES	LOW	N/A
I4-Control*	YES	YES	NO	YES	LOW	N/A
Head stick**	YES	YES	NO	NO	HIGH	>100USD
Jouse**	NO	NO	YES	NO	MEDIUM	>100USD
Tongue Control**	YES	YES	YES	YES	LOW	N/A
Nessi***	YES	YES	NO	YES	MEDIUM	N/A
BCI – SSVEP***	YES	YES	NO	YES	MEDIUM	>100USD

*) – vision-based interface, **) – mechanical rigs, ***) – EEG recording.

II. HEAD-TRACKING HUMAN-COMPUTER INTERFACE

In the author's opinion, vision-based HCI solutions feature many advantages. They are both contact-free and non invasive. Such systems are simple in use, require no force, no mechanical contact a little maintenance. What is more, the vision-based systems can be "always on" – ready for use and awaiting a user's commands. Finally, they are medically safe and ethically easy-acceptable.

The proposed vision-based human-computer interface system consists of two major building blocks [10]:

- hardware – providing acquisition image sequences,
- software – containing algorithms for image analysis, screen cursor displacement calculation and program control; the software is implemented with the use of the OpenCV library.

The system bench is illustrated in Fig. 1. The user is positioned in front of the computer screen at a distance within the range of 20-50 cm. The camera can be a built-in device or attached to the monitor.

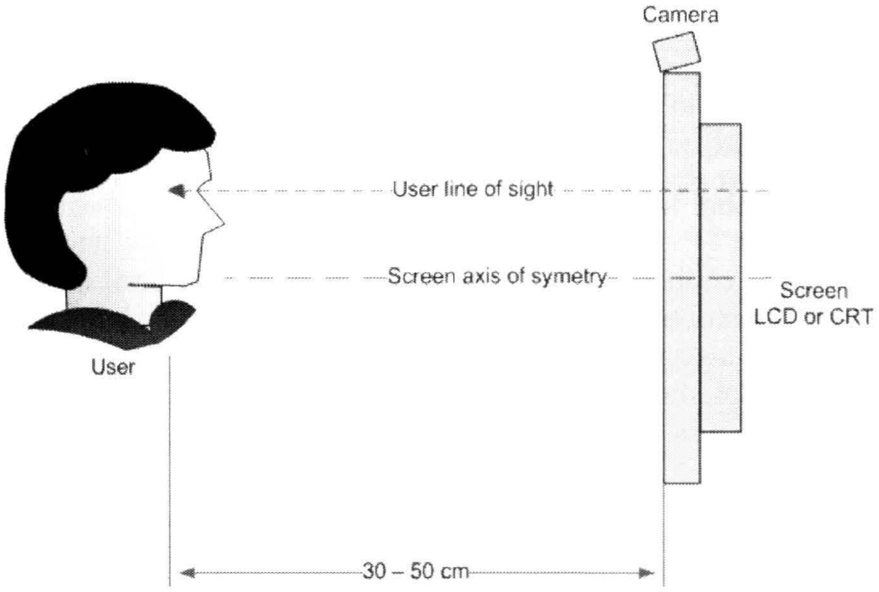


Figure 1. System user positioning with respect to computer screen and the camera

A. *Software design and implementation*

A PC Windows OS application for this project is to perform the following operations:

- capture a sequence of images of the user's face,
- preprocess the images,
- identify the user's face location in consecutive image frames,
- resolve current face movements and decide of mouse cursor shift,
- detect the user's eyes (open/closed) and perform specified mouse button actions**.**

In order to perform the indicated operations an application was designed, consisting of the software modules that are indicated in a block diagram shown in Fig. 2.

The image acquisition and pre-processing modules fetch frames from the imager and perform the following image pre-processing operations: resolution adjustment, image rotation, image flipping, color coding transformation and others.

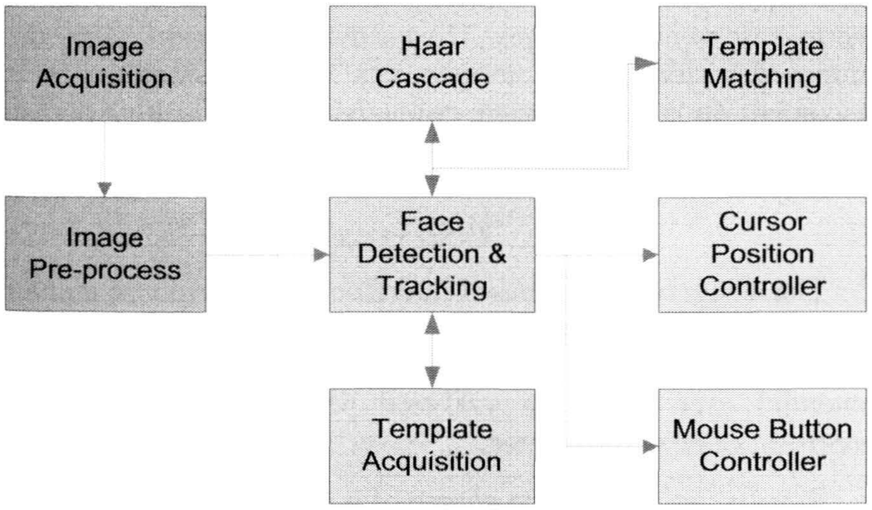


Figure 2. Software modules implemented for image-based head tracking and eye-blink detection

The face detection and tracking module consists of the three image processing steps: 1) face detection: by means of Haar cascade filtering masks according to Viola and Jones algorithm [11], 2) tracking initialization: for the motionless "neutral" face position, face and eye templates are collected, 3) face tracking: implemented by means of template matching technique which serves both for face tracking and monitoring the user's blinking.

During tracking each frame retrieved from the camera is correlated with the face template. The point with the highest match is assumed as the face's geometric center of mass and its coordinates are forwarded to the Cursor Position Controller (Fig. 3).

The Mouse Button Controller is activated in a "neutral" enface face position. The eye template acquired during initialization is assumed to represent an "eyes open" state. Once the "quality" of the match falls below a configurable threshold, the left eye is identified as closed. This triggers the left mouse button service routine. If the user's face leaves the neutral position while the left eye is identified as closed dragging or framing is performed. Screen cursor movements are controlled by user's head pitching and yawing (Fig. 3).

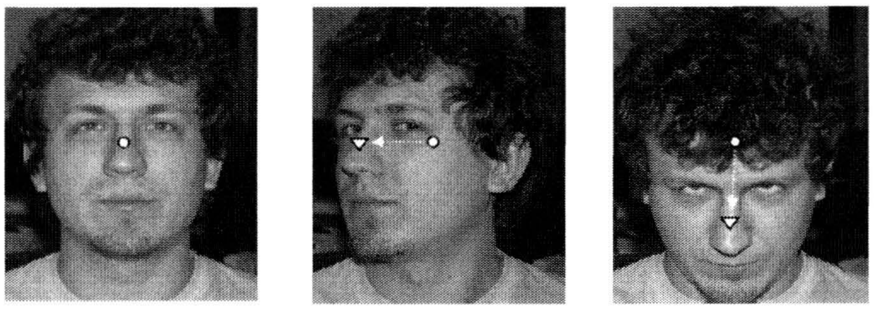


Figure 3. Image of user's face from the left: "neutral" enface position, head yawn, head pitched; circular marker represents center of the face in the "neutral" position; triangular marker represents center of face after rotation, its displacement from "neutral" position is clearly visible

III. System Tests and Results

Hardware setup of the system is bound to be simple and limited to commercially available, off-the-shelf components. There are no specific requirements for the camera's resolution, though it is assumed to be at least VGA standard resolution of 640x480 pixels with a frame rate between 25 and 30 fps (frames per second). For image acquisition an ordinary webcam with USB interface was used: Logitech model QuickCam UltraVision. As a computing platform for the software a PC class computer with an Intel Celeron 1,5 GHz processor and 512 MB RAM memory was used and proved to be sufficient for running the application at a pace set by the camera frame rates.

The system was tested in 4 trial scenarios on 6 test subjects. Test users were both frequent computer users, accustomed to Microsoft Windows XP operation, working with a computer several hours each day and "occasional" users who in general lacked this experience. Trials had been preceded with individual training sessions.

Each trial session consisted of three parts that were conducted in different lighting conditions:

1. Optimal lighting conditions, in midday, with good natural illumination.
2. Poor conditions, in late evening hours, with artificial light sources placed outside the camera's field of view.
3. Complex lighting conditions with external light source placed in the field of view, behind and above the user.

Four test scenarios were prepared to simulate situations experienced in typical computer use situations. Scenarios were focused on leisure and entertainment. The lack of keyboard input precludes the vision-based interface in the current version from being used as a tool for work

A. Test scenario 1 – web browsing

This scenario simulates use of the interface for accessing information using a popular news portal. It is assumed that the necessary bookmarks have been added to the web browser.

In this scenario the user is expected to perform the following actions: 1) Open the Opera web browser using the shortcut on the workspace, 2) Open the Bookmarks tab, 3) Select one of 5 available bookmarks, 4) Access the weather forecast section using an expandable menu, 5) Return to the root bookmark, 6) Access the news section, 7) Scroll the page down, 8) Select the last available news article.

Total time and number of errors is recorded. Errors are divided into following types: I. Involuntary actions – e.g. an unintended click, II. Time related errors – e.g. a single click instead of a double, III. Precision related errors – e.g. failing to use the expandable menu.

The total average time to complete all targets in this scenario was less than 5 minutes. This speed seems to be poor in comparison to manually operated input devices. Nevertheless, error of type I –has never occurred in this scenario. Less severe errors of type II and III – related with incorrect timing when clicking and lack of precision in navigating, had no significant impact on the final result.

B. Test scenario 2 – image browsing with IrfanView

Simulates accessing and browsing a collection of 10 stored images. The user is to perform the following actions: 1) Double click on My Computer, 2) Access drive C:\, 3) Access a specified folder, 4) Access the folder Pictures, 5) Open the first picture, 6) Scroll through all 10 pictures using "Next file..." button, 7) Close IrfanView.

All users succeeded in completing all tasks. The average time of 1 minute and 40 seconds to complete all goals is a promising result, comparable to manually operated input devices. The source of this improvement is the spatial distribution of the used icons and buttons. Comparing to conditions of scenario 1, where the hyperlinks were spread on the whole screen IrfanView requires the use of only the toolbar where all buttons are located in close vicinity. This considerably decreased the total path traversed by the cursor.

C. Test scenario 3 – the labyrinth

This scenario is focused on precision of cursor movement. The goal is to lead the pointer through a labyrinth without touching the edges (see Fig. 4). The smallest width of the path is set to 10 pixels. The total size of the labyrinth is 700x500 pixels.

In the course of this scenario precision and control of the cursor position was tested. In Fig. 4 a typical labyrinth following task is illustrated (all subjects obtained comparable results). Satisfactory control in vertical and horizontal cursor movement was noted. Traversing a skewed line proved a more difficult task.

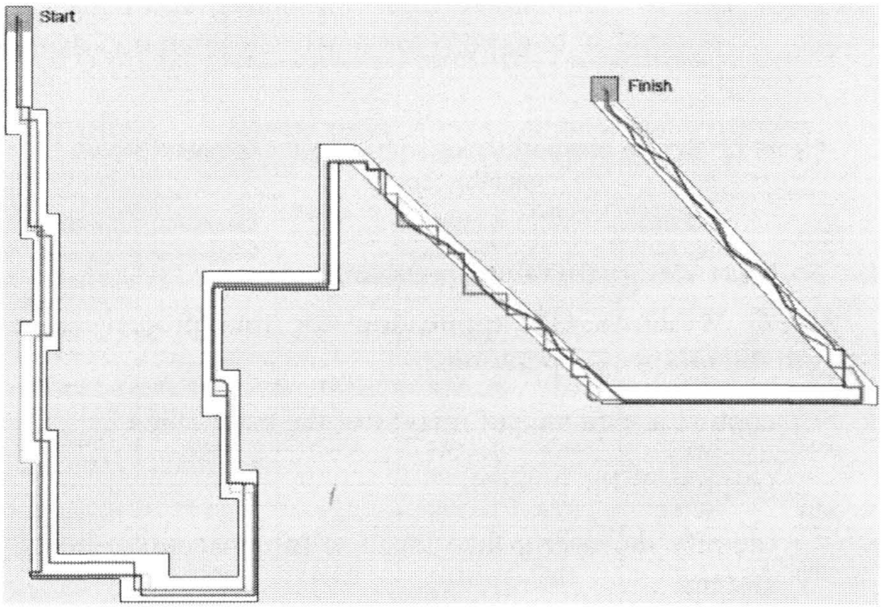


Figure 4. Paths traversed in the labyrinth by subject 1; the lines indicate results from three conducted sessions

D. Test scenario 4 – reading pdf files with Adobe Reader

This scenario simulates opening and manipulating a pdf format file using Adobe Reader. The following actions are to be performed: 1) Open Start expandable menu, 2) Select a folder, 3) Access a folder, 4) Open test.pdf, 5) Zoom in, 6) Scroll down to second page continuously, 7) Scroll, 8) Close Adobe Reader.

Average time to complete all goals was 1 minute and 53 seconds. A user when reading a book spends most of his/her time focused on the text. Mouse actions are limited to scrolling the text or switching pages. Thus a disabled person, using the proposed vision based interface, can read a voluminous text document in a time period comparable to a healthy person using mouse or touchpad interfaces.

IV. CONCLUSIONS

The designed vision-based interface provides mouse pointer control and allows performing single and double left mouse button clicks using head movements and eye blinks. A potential user can be a paralyzed person with deteriorated precision of head movement.

Results gathered in the course of 4 test scenarios prove that the presented prototype meets the design constrains and requirements. Implemented solutions demonstrate the necessary precision and reliability. Although, this vision-based interface features significantly reduced input speed, as compared to manually operated devices, it is still a promising step forward for making computers accessible to physically disabled persons. Such perspective is a significant improvement in life quality, providing a sense of independence and self-sufficiency.

References

[1] A. Helal, M. Mounir, and B. Abdulrazak, B. (Eds.), The Engineering Handbook of Smart Technology for Aging, Disability, and Independence, John Wiley & Sons, Inc., 2008.

[1] EyeTech Digital systems web-page, www.eyetechds.com/ (accessed on 4.02.11).

[2] A. Królak, P. Strumiłło, Eye-blink detection system for human computer interaction, *International Journal on Universal Access in the Information Society*, vol. 10, 2011, DOI 10.1007/s10209-011-0256-6.

[2] I4Control® System webpage: http://cyber.felk.cvut.cz/ic4/ (accessed on 4.02.11).

[3] Forbes Rehab Services web page: www.frs-solutions.com (accessed on 4.02.11).

[4] Compusult Services web page: www.jouse.com (accessed on 4.02.11).

[5] L. N. Struijk, E.R. Lontis, B. Bentsen, H.V. Christensen, H.A. Caltenco, and M.E. Lund, "Fully integrated wireless inductive tongue computer interface for disabled people", Conference Proceeding IEEE Engineering Medical Biological Society, pp. 547 550, 2009.

[6] M. Bensch, A. Karim, J. Mellinger, T. Hinterberger, M. Tangermann, M. Bogdan, W. Rosenstiel, and N. Birbaumer, "Nessi: An EEG-Controlled Web Browser for Severely Paralyzed Patients", Computational Intelligence and Neuroscience, 2007.

[7] A. Materka, M. Byczuk, "Alternate half-field stimulation technique for SSVEP-based brain-computer interfaces", Electronics Letters, vol. 42, no. 6, pp. 321 322, 2006.

[8] T. Pajor, Design of a Vision-Based Human-Computer Interface for the Physically Disabled, MSc Thesis, Technical University of Lodz, 2010.

[9] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, pp. I 511 I 518, 2001.

Cluster Analysis of Canonical Correlation Coefficients for the SSVEP Based Brain-Computer Interfaces

Paweł Poryzała
Institute of Electronics,
Technical University of Lodz
Lodz, Poland
e-mail: poryzala@p.lodz.pl

Andrzej Materka
Institute of Electronics
Technical University of Lodz
Lodz, Poland
e-mail: materka@p.lodz.pl

Marcin Byczuk
Institute of Electronics
Technical University of Lodz
Lodz, Poland
e-mail: byczuk@p.lodz.pl

Abstract — **In this paper, a novel method for detecting steadystate visual evoked potentials (SSVEP) using multiple channel electroencephalogram (EEG) data is presented. Accurate asynchronous detection, high speed and high information transfer rate can be achieved after a short calibration session. Spatial filtering based on the Canonical Corelation Analysis method proposed in [1] is used for identifying optimal combinations of electrode signals that cancel strong interference signals in the EEG. Data from a test group consisting of 21 subjects are used to evaluate the new methods and to compare results to standard spectrum analysis approach. Conducted research, for different length signal segments and five visual frequencies, showed improvement of both classification accuracy and detection speed.**

Index Terms — Brain-Computer Interface (BCI), Electroencephalogram (EEG), Steady-State Visual Evoked Potential (SSVEP) detection

I. Introduction

Studies on the development of the Brain-Computer Interfaces (communication systems, that do not depend on the brain's normal output pathways of peripheral nerves and muscles [2]) have more than 20-year history. BCI devices may allow people with disabilities, including paralysed people, use the computer and other technical equipment, on a par with other users. Over the years, most widely represented group of devices are non-invasive BCI systems with electroencephalographic (EEG) brain activity monitoring.

At the moment, the most commonly used EEG-based BCI systems employ event-related synchronization of μ and β rhythms (ERD/ERS), event-related potentials (ERP) and steady-state visual evoked potentials (SSVEP). Information transfer rate (ITR, introduced in [3]) is used by majority of the BCI laboratories and research groups to evaluate BCI system performances. This measure depends on three factors: speed, accuracy and number of targets. It is proved, that currently the SSVEP approach provides the fastest and the most reliable communication paradigm for the implementation of a non-invasive BCI system [4].

High speed and accuracy, sufficient number of targets for a particular task are essential for BCI system in order to become a practical device. Today a number of signal processing methods for detection and extraction of SSVEPs exist. From simple methods for detecting a single frequency component in a single electrode signal [5], through most widely used spectrum analysis methods [6], [7] up to multichannel spatial filtering and detection methods [8], [1].

In this paper, a novel approach for multichannel detection of SSVEP responses is proposed. System, after a simple calibration session, is able to work asynchronously with improved (in relation to spectrum analysis method) detection speed and accuracy (thus higher ITR).

The paper is organized as follows. The second section discusses the details of the proposed method. Off-line experiment conducted to prove the algorithm quality are presented in the third session. Fourth section contains results and discussion. Conclusions are presented in the last section.

II. Detection Method

In this section, the proposed Cluster Analysis Canonical Correlation (CACC) method for detection of SSVEPs is discussed. It is based on the coefficients derived from the Canonical Correlation Analysis (CCA) which is described in what follows.

A. Canonical Correlation Analysis

CCA method is used for finding the correlations between two sets of multi-dimensional variables. It was first used for SSVEP detection in [1] and was further developed in [9].

CCA method seeks for a pair of linear combinations **w** and **v**, for two sets of data **Y** and **X**, such that the correlation

$$\rho_1 = \mathrm{cor}(\mathbf{S}, \mathbf{U}) \qquad (1)$$

between the first pair of canonical variables $\mathbf{S} = \mathbf{w}^T\mathbf{Y}$ and $\mathbf{U} = \mathbf{v}^T\mathbf{X}$ is maximized. Consecutive pairs of linear combinations, canonical variables and canonical correlation coefficients can be obtained, but the maximum number of pairs equals the number of variables in the smallest of two sets (**Y** and **X**).

As far as the CCA method is used for SSVEP detection: **Y** refers to the set of N_y multi-channel EEG signals and **X**

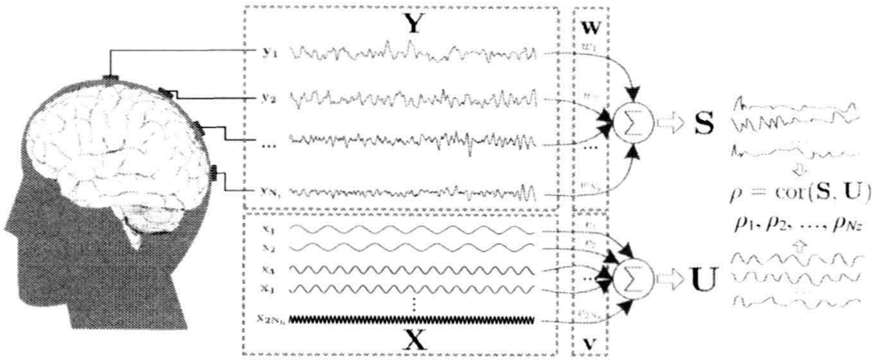

Y
W
y_1
y_2
...
y_{N_y}
S
$\rho = \mathrm{cor}(\mathbf{S}, \mathbf{U})$
$\rho_1, \rho_2, \ldots, \rho_{N_f}$
x_1
x_2
x_3
x_4
x_{2N_h}
U
X
V

Fig. 1: An illustration for usage of the CCA method in EEG signal analysis. Matrix $\mathbf{Y}$ is where data from N_y EEG channels is stored. $\mathbf{X}$ is an ideal, reference SSVEP response, containing both sinus and cosinus components for N_h harmonics.

refers to the set of $2N_h$ reference signals (Fig. 1). In the rows of the $\mathbf{X}$ reference matrix, the sinus and cosinus components for all N_h harmonics of the stimulation frequency are stored:

$$\mathbf{X} = \begin{bmatrix} \sin(2\pi f t) \\ \cos(2\pi f t) \\ \ldots \\ \sin(2\pi N_h f t) \\ \cos(2\pi N_h f t) \end{bmatrix}. \tag{2}$$

CCA finds the maximum canonical correlation with respect to weight vectors $\mathbf{w}$ and $\mathbf{v}$ by solving the following problem:

$$\begin{aligned} \max_{\mathbf{w},\mathbf{v}} \rho &= \frac{\mathrm{cov}[\mathbf{S},\mathbf{U}]}{\sqrt{\mathrm{var}[\mathbf{S}]\,\mathrm{var}[\mathbf{U}]}} \\ &= \frac{E[\mathbf{SU}]}{\sqrt{E[\mathbf{S}^2]E[\mathbf{U}^2]}} \\ &= \frac{E[\mathbf{w}^T\mathbf{Y}\mathbf{X}^T\mathbf{v}]}{\sqrt{E[\mathbf{w}^T\mathbf{Y}\mathbf{Y}^T\mathbf{w}]E[\mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}]}}. \end{aligned} \tag{3}$$

When CCA is used in frequency recognition of the SSVEP-based BCI system, where there are N_f targets (stimulus frequencies $f_1, f_2, \ldots, f_{N_f}$), the same number of reference matrices must be used (Fig. 2).

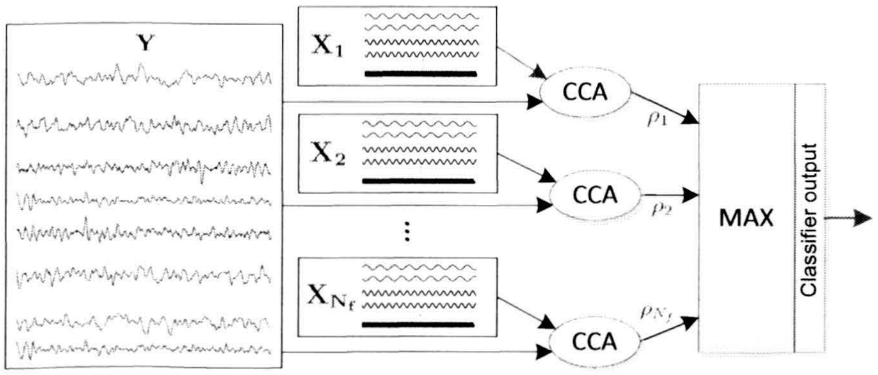


Fig. 2: An illustration for usage of the CCA method in different frequency components recognition of the SSVEP-based BCI where there are N_f targets. $\mathbf{X_i}$ is the response reference matrix for the i-th stimulus frequency.

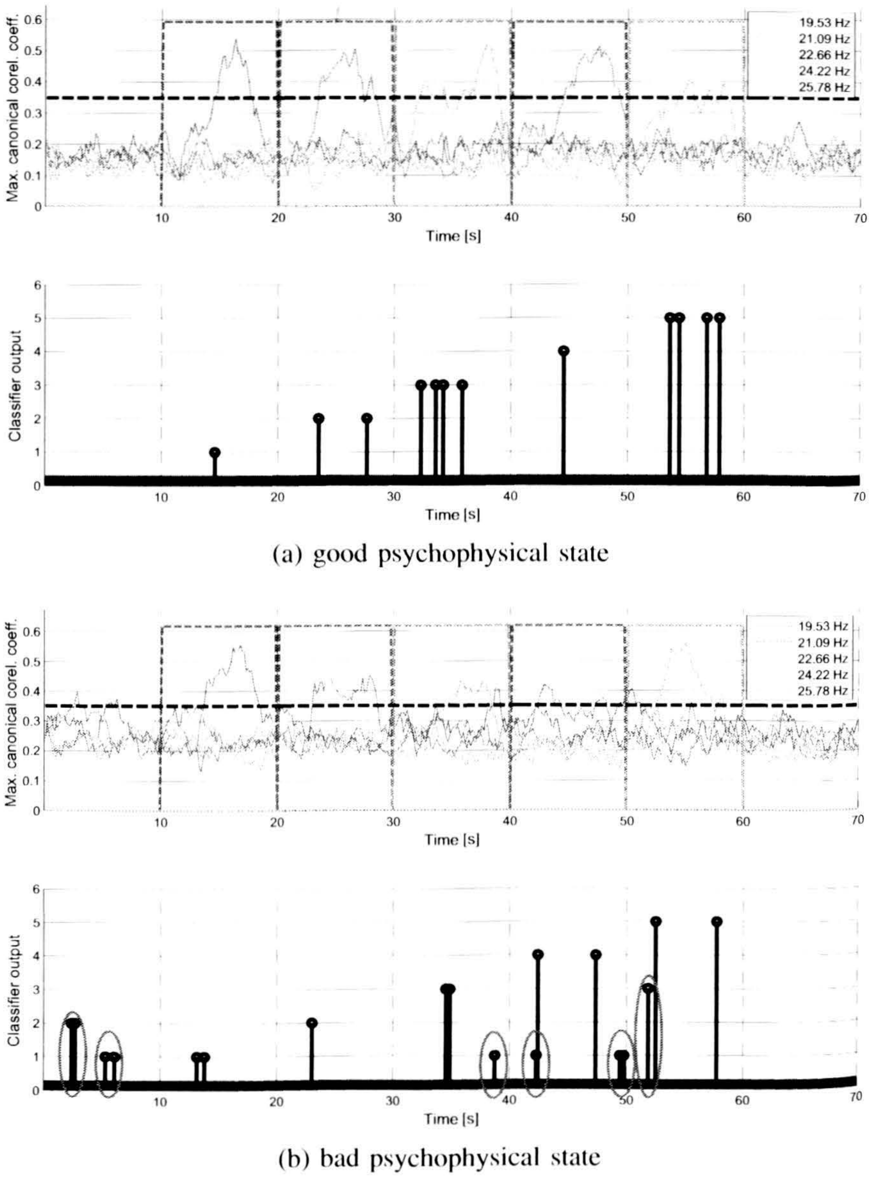
For each pair of multi-channel EEG and reference signals, a maximum canonical correlation coefficient is obtained and it can be used for frequency recognition. As proposed in [9] user's command is recognized as

$$C = \max_i \rho_i, \quad i = 1, 2, \ldots, N_f, \tag{4}$$

where ρ_i is the CCA coefficient obtained with the reference signal frequency being $f_1, f_2, \ldots, f_{N_h}$.

B. Encountered CCA problems

Original CCA method, even in conjunction with thresholding of maximum canonical correlation coefficients for each stimulus frequency, does not seem reliable in practical, asynchronous SSVEP BCI system. Main problem is related to strong dependence of measured EEG signals against user psychophysical state (Fig. 3). This state changes with the ongoing measurement session and between different days (when user eg. did not sleep well). In such cases, the background, non-stimulated EEG activity is increased. Brief moment of relaxation often improves recorded signal quality, but this is usually only a short-term effect.



(a) good psychophysical state

(b) bad psychophysical state

Fig. 3: Comparison of the classification results of the relevant parts of the signal for the AL1 user in two test sessions. In the second case wrong system decisions were marked red.

In Fig. 3b it is clearly visible that all of the canonical correlation coefficients have greater variability over time (often reaching established threshold value, resulting in false detections). In this particular example SSVEP BCI system is not able to distinguish between working and idle state classes properly. There is also only a little margin to rise threshold value due to the low canonical correlation coefficient values in segments which involved stimulation.

C. Cluster Analysis Canonical Correlation

Original CCA method uses a single canonical correlation coefficient (with the highest value) for each of the N_f SSVEP response patterns. CACC method uses three highest valued correlation coefficients as features. Detection and idle states can be accurately identified with k-means cluster analysis performed separately in each of the feature spaces (Fig. 4).

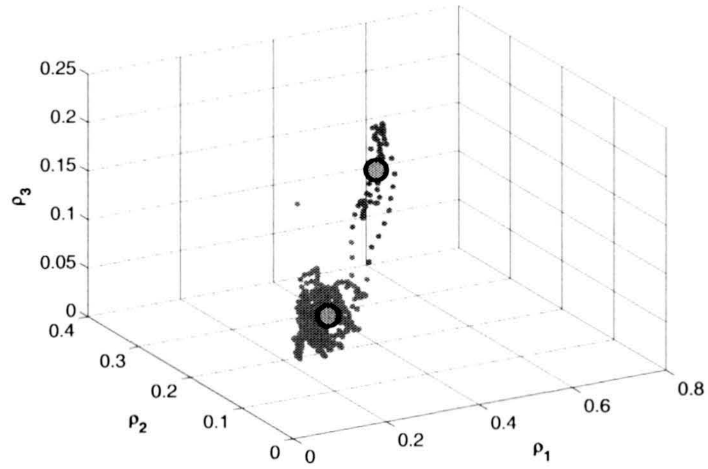


Fig. 4: Sample result of the k-means cluster analysis in the correlation coefficients feature space. Idle class was marked red, detection class was marked blue. Centroids of both classes were marked green.

Distance which can be measured between centroids of both detection and idle classes in feature spaces for each stimulus frequency, varies between the subject and frequency used for stimulation. Its value must be determined during the training phase, therefore BCI system work should be divided into two stages:

1) calibration session: At this stage (Fig. 5) the objective of the algorithm is to identify the distances between centroids of detection and idle classes. This value is characteristic for each of the frequencies used for stimulation. The user is instructed to move his/her eyes (but not faster than every 5 seconds) between all stimulation symbols.

In the first step, a set of response patterns for each of the stimulation frequencies used ($\mathbf{X_i}, i = 1, 2, ..., N_f$) is built. As a result of canonical correlation of N_y EEG source channels in the detector window $\mathbf{Y}$ sequentially with patterns $\mathbf{X_i}$, one gets sets of three factors: ρ_{1i}, ρ_{2i} and ρ_{3i}. Each of the sets can be represented as a point ρ_i in the the feature space constructed on the basis of canonical correlation coefficients of the source EEG data with the i-th response pattern.

Along with each successive point ρ_i in particular feature space, k-means cluster analysis is performed and mutual distance between two classes (detection and idle state) is examined. Euclidean metric is used:

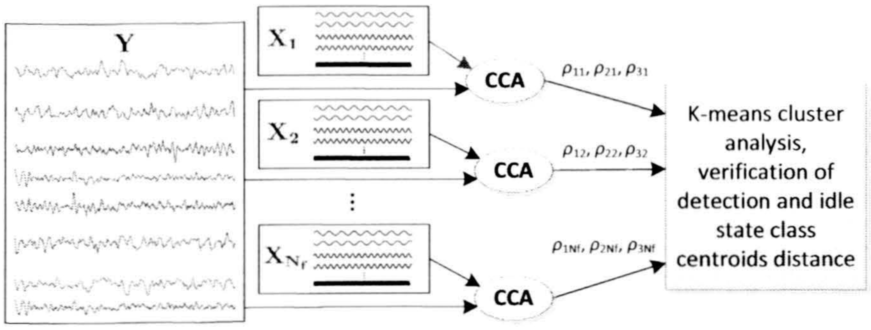

Y
$\mathbf{X_1}$
$\mathbf{X_2}$
⋮
$\mathbf{X_{N_f}}$
CCA
CCA
CCA
$\rho_{11}, \rho_{21}, \rho_{31}$
$\rho_{12}, \rho_{22}, \rho_{32}$
$\rho_{1Nf}, \rho_{2Nf}, \rho_{3Nf}$
K-means cluster analysis, verification of detection and idle state class centroids distance

Fig. 5: System in calibration mode.

$$d(B_i, D_i) = \sqrt{\sum_{j=1}^{3}(\rho_{B,ji} - \rho_{D,ji})^2}, \tag{5}$$

where B_i and D_i denote the points where the idle and detection class centroids lay in the i-th feature space. Calibration of the frequency f_i ends when the B_i and D_i centroid distance is large enough:

$$d(B_i, D_i) \geq \beta \tag{6}$$

and after adding e.g. the last 25 points to appropriate feature space, the distance was not changed by more than 10%.

Based on the analysis of recorded EEG data and our practical investigations, $\beta = 0.25$. Its value is a compromise between the accuracy (especially for lower quality signals) and the time of detection. Too high β value results in increased number of false negative errors, and too small increases false positives.

The training session ends upon completion of the calibration for all N_f frequencies. If the calibration procedure lasts over one minute, the system reports a problem with particular frequency.

2) working mode: This is the target operating mode, in which device is used for communication (Fig. 6). All calibrated data (locations of the detection and idle class centroids in each of N_f feature spaces) are used to improve classification at this stage.

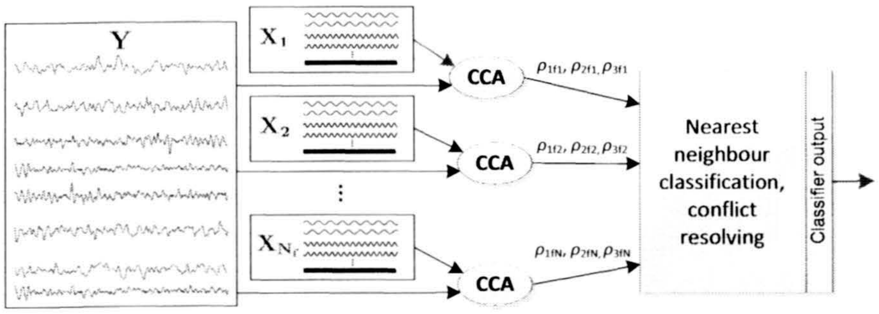

Y
$\mathbf{X_1}$
$\mathbf{X_2}$
⋮
$\mathbf{X_{N_f}}$
CCA
CCA
CCA
$\rho_{1f1}, \rho_{2f1}, \rho_{3f1}$
$\rho_{1f2}, \rho_{2f2}, \rho_{3f2}$
$\rho_{1fN}, \rho_{2fN}, \rho_{3fN}$
Nearest neighbour classification, conflict resolving
Classifier output

Fig. 6: System in working mode.

Like in the calibration mode, as a result of canonical correlation of the EEG source data ($\mathbf{Y}$) with subsequent response patterns X_i, sets of three coefficients: ρ_{1i}, ρ_{2i} and ρ_{3i} (a point ρ_i in a three dimensional feature space) are obtained. Each point is classified (nearest neighbours method) to one of the classes B_i or D_i.

If, during the classification in each of N_f feature spaces, none or exactly one point ρ_i was classified to D_i class, system will detect respectively class zero (idle state) or number i of particular feature space. If more than one point, represented by the canonical correlation analysis coefficients of source data and response pattern X_i, will be qualified to the detection classes, a conflict occurs.

Conflict situations are solved by using the distance of each of the conflicted points ρ_i from the point laying on the line passing through centroids of both B_i and D_i classes, and lying half-way between them. The classifier output is determined as the number of the i-th feature space in which the distance was the greatest.

After successful detection of responses at any of the stimulus frequencies, all data in detector window $\mathbf{Y}$ are replaced with zeros. This prevents multiple detection of the same symbol. In addition, after each classification, 700 ms of the EEG data will not be utilized (classifier will not take any decisions). This will give the user of the BCI system time for gaze shifting.

III. Offline experiments

The experiments were carried out at the Institute of Electronics, Technical University of Lodz. Fig. 7 presents the layout of the measurement stand.

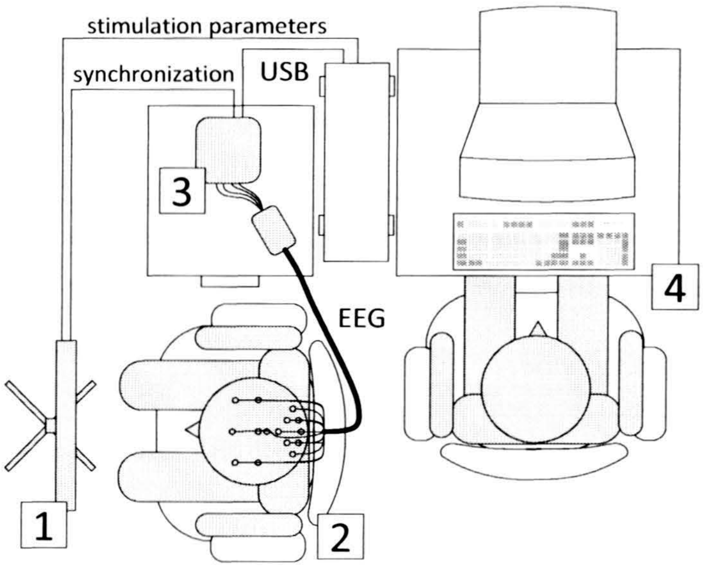


Fig. 7: Layout of the measurement stand: visual stimulator (1), subject (2), EEG recording device (3) and operator (4).

Subjects sat in the front of a visual stimulator (described in the next section) on a comfortable, ergonomic chair. Measurements were carried out in a room with a window on the south side, curtained with a light impermeable material blind and a standard fluorescent light switched on. Light conditions during all experiments were the same.

A. Subjects

Twenty one healthy subjects (ten women and eleven men, age range 16-33 years, with the average of 22.2 years and a standard deviation of 3.4 years) participated in this study. For each subject, two measurements were carried out on different days.

Four subjects previously used our BCI system. None of the subjects had neurological or visual disorders (glasses or contact lenses were worn where appropriate). Subjects did not receive any financial rewards.

In the early stages of the experiment, users were qualified to one of three groups:

1) Group A (best results, 5 subjects): Subjects who in most cases had earlier contact with the device (in our previous studies and tests).

2) Group B (average results, 11 subjects): The most widely represented group. Subjects who were not familiar with the idea of a BCI device, but actively participated in the experiments.

3) Group C (poor results, 6 subjects): Subjects with concentration problems or very high unstimulated, spontaneous brain activity

This classification helped to investigate system parameters in relation to a specific group of users.

B. Visual Stimulator

A universal, computer driven LED stimulator was used for stimulation. Each stimulation symbol (Fig. 8) consisted of three LEDs: two stimulation lights with a diameter of 5mm positioned on the lower right and lower left quarter of the visual field of each eye retina and one fixation light with a diameter of 3mm placed in the center of visual field. Distance from visual stimulator to subject eyes was equal to 50 centimetres.

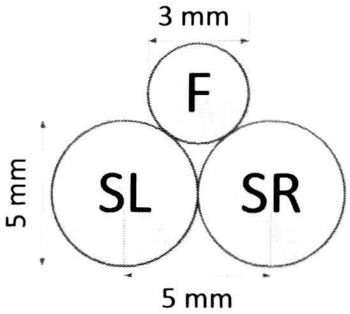


Fig. 8: A view of stimulating lights (SL, SR) and a fixation light (F) on the screen of stimulator.

Stimulation lights flash with the same frequency alternatively in phase (alternate half-field stimulation technique [10]). Fixation light is used for two purposes: the subject is expected to concentrate his/her sight on it; additionally it provides a feedback information about amplitudes of corresponding SSVEPs detected in the subject EEG signal.

Visual stimulator had five sets of LEDs forming stimulation symbols in five different colors (each set had stimulation and fixation LEDs of the same color): white, blue, green, yellow and red. Luminous intensity of each LED used was approximately 1000mcd.

C. EEG Recording

Equipment from g.tec (Graz, Austria) was used for EEG measurements: g.USBamp biosignal amplifier, g.GAMMAbox active electrode driver and g.GAMMAcap with sixteen

Ag/AgCl active electrodes. Seven electrodes over the primary visual cortex (positions PO7, PO3, O1, OZ, O2, PO4 and PO8) and nine electrodes evenly distributed over the remaining cerebral cortex (positions P3, PZ, P4, C3, CZ, C4, F3, FZ and F4) were used for recording. A ground electrode was placed on CPZ position. A reference electrode was placed on right ear lobe (position A2). The EEG signals were bandpass filtered between 2.0-60.0 Hz with a notch filter for 50 Hz power line frequency suppression, amplified and sampled at 600 Hz.

EEG signals were recorded with a home-made software package - BioStudio [11] which was able to drive visual stimulator and processed measured signals in order to compute biofeedback information for stimulation symbols.

D. Experimental paradigm

Subjects were instructed to focus their gaze on fixation LED and flickering lights below it to produce SSVEPs. Each measurement lasted for several minutes and consisted of five stimulus sequences (one sequence for each color, only one stimulation symbol switched on at a time). The first sequence began a few seconds after starting the measurement (time required for stabilization of electrode-skin connection impedance and possible adjustments of subjects' position on the chair to reduce the EMG signals). Stimulation frequencies were chosen to match the discrete Fourier transform frequencies used in the subsequent analysis (in order to minimize spectral leakage). Each sequence contained 27 different stimulation frequencies in the range of about 7–47 Hz.

Each stimulation lasted eight seconds, followed by a 2-second pause before the next stimulation (Fig. 9). Additionally a brief pause followed each sequence (several up to tens of seconds). This pause was intended for position adjustments on the chair and subject relax with eyes closed (EEG signal was still being recorded).

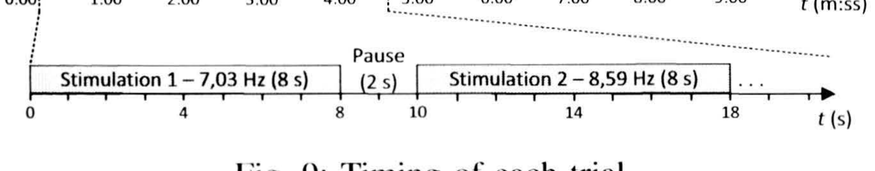

manually controlled pause for relax and position adjustments
automatic sequence generation
automatic sequence generation
Sequence 1 – white (4:30 min)
Sequence 2 – blue (4:30 min)
. . .
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
t (m:ss)
Pause
(2 s)
Stimulation 1 – 7,03 Hz (8 s)
Stimulation 2 – 8,59 Hz (8 s)
. . .
0
4
8
10
14
18
t (s)

Fig. 9: Timing of each trial.

Binary signal from visual stimulator indicating stimulation state (on/off) was recorded along with the subject EEG signal from all sixteen channels.

The original EEG data for subjects were re-sampled ($F_s = 200$ Hz) and divided into shorter fragments, containing several stimulation patterns. Algorithm was tested with window lengths of 1.28, 2.56 and 5.12 seconds and data window moved with a step of 0.16 s.

Results of the proposed method were compared to standard spectrum analysis SSVEP detection approach: power spectral density of EEG signal was computed in the sliding window (frequency resolution of about 0.78 Hz). For each predefined discrete frequency of stimulation a signal to background ratio (SBR) was estimated [12]. The frequency of the maximum SBR, after it was compared with the threshold value, was decided to be the intended target of the user. This algorithm was executed for all possible bipolar source electrode combinations:

$$C_{N_y}^2 = \binom{N_y}{2} = \frac{N_y!}{2!(N_y - 2)!} \tag{7}$$

in order to find Best Bipolar Combination (BBC). In analysed case ($N_y = 16$) 120 bipolar channels had to be processed.

TABLE I: Results in Group A

Window Length [s]	accuracy [%]		det. speed [s]		ITR [bpm]	
	BBC	CACC	BBC	CACC	BBC	CACC
1.28	91.19	90.27	2.55	2.28	40.38	43.82
2.56	94.12	93.05	2.47	2.52	45.74	43.29
5.12	91.53	94.88	4.02	3.35	25.85	34.52

IV. Results and discussion

Binary markers (stimulation on and off events) stored in parallel with the EEG data and the known stimulation sequence for each color were used to verify performance of the proposed detection algorithm. Classification results for each user were assessed in terms of accuracy, average detection time and the information transfer rate and were afterwards averaged in each of the subject groups.

A. Group A

High accuracy of both SSVEP detection methods is proved (Table I). The increase in detection accuracy with the increase of window length is negligible. Measured mean detection times increase as the window length is extended (this is a known problem an can be easily solved in practical system by use of multiple, different length parallel detectors). Information transfer rates are similar in case of both algorithms.

B. Group B

The biggest increase of accuracy of the CACC method over the BBC algorithm was observed in this group (Table II). Depending on the window length, it was from 7 up to 11%. There is also a noticeable rise of detection accuracy with the increase of window length. As in the previous group, average detection times are similar (particularly for shorter windows), but detection usually took about 0.8–1.0 second longer. As far as the information transfer rate is considered, CACC method seams to be clearly better than BBC because of both: shorter detection times and higher accuracy.

C. Group C

Increase of detection accuracy for the CACC method over the BBC algorithm in this group was from 5 up to 8% (Table III). Similarly to the first group, the increase in detection accuracy with the increase of window length is negligible. As far as the average detection speed is considered, BBC method

TABLE II: Results in Group B

Window Length [s]	accuracy [%]		det. speed [s]		ITR [bpm]	
	BBC	CACC	BBC	CACC	BBC	CACC
1.28	63.15	70.74	3.39	3.02	11.24	17.18
2.56	68.26	79.65	3.26	3.15	14.46	22.58
5.12	68.33	78.51	5.06	4.15	9.34	16.51

TABLE III: Results in Group C

Window Length [s]	accuracy [%]		det. speed [s]		ITR [bpm]	
	BBC	CACC	BBC	CACC	BBC	CACC
1.28	45.07	50.98	4.02	5.06	3.44	2.14
2.56	47.23	53.22	4.75	5.12	3.39	4.06
5.12	47.12	55.17	5.35	5.72	2.99	4.54

is faster (difference of about 1 s for the shortest window and about 0.4 s in remaining cases).

V. Conclusions

Results clearly show that research on multichannel detection methods are important and can significantly improve classification accuracy, detection times and overall communication speed. The proposed detection method improves the classification accuracy in the groups of subjects with the average (Group B) and poor (Group C) results. In the group of users with the best results (Group A), there was no clear improvement of the SSVEP detection accuracy. Average detection times for both algorithms are similar in most cases (but there were differences of up to 1 second). Information transfer rate in many cases (especially for Groups B and C) was higher for the CACC method, which is due to greater classification accuracy of this method. What is important, only a short off-line calibration session was necessary to achieve such results.

At the moment many of the BCI systems are at the stage of laboratory demonstrations. This is mainly due to high user variation, BCI illiteracy phenomenon and low communication speeds (low ITR). New spatial filtering and detection methods will make it possible to overcome this limitations. In the presented research, each of 21 subjects was able to communicate in the off-line experiments and 16 subjects (Groups A and B) reached substantial information transfer rates. These results encourage further development of the proposed detection method and its implementation in the online BCI system, what will be the subject of our future work.

Acknowledgment

This work is partially supported by Polish Ministry for Science and Higher Education grant NN515 520838.

References

[1] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency Recognition Based on Canonical Correlation Analysis for SSVEP-Based BCIs," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 6, pp. 1172–1176, Jun. 2007.

[2] J. Wolpaw, N. Birbaumer, W. Heetderks, D. McFarland, P. Peckham, G. Schalk, E. Donchin, L. Quatrano, C. Robinson, and T. Vaughan, "Brain-computer interface technology: a review of the first international meeting," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 164–173, Jun. 2000.

[3] J. Wolpaw, H. Ramoser, D. McFarland, and G. Pfurtscheller, "EEG-based communication: improved accuracy by response verification," *IEEE Trans. Rehabil. Eng.*, vol. 6, no. 3, pp. 326–333, Sep. 1998.

[4] Y. Wang, X. Gao, B. Hong, C. Jia, and S. Gao, "Brain-Computer Interfaces Based on Visual Evoked Potentials," *IEEE Eng. Med. Biol. Mag.*, vol. 27, no. 5, pp. 64–71, Sep. 2008.

[5] J. S. Victor and J. Mast, "A new statistic for steady-state evoked potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 78, no. 5, pp. 378–388, May 1991.

[6] Y. Wang, R. Wang, X. Gao, B. Hong, and S. Gao, "A practical VEP-based brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 234–240, Jun. 2006.

[7] A. Materka, M. Byczuk, and P. Poryzala, "A Virtual Keypad Based on Alternate Half-Field Stimulated Visual Evoked Potentials," in *Int. Symp. on Inf. Technol. Converg., ISITC 2007*, 2007, pp. 296–300.

[8] O. Friman, I. Volosyak, and A. Graser, "Multiple Channel Detection of Steady-State Visual Evoked Potentials for Brain-Computer Interfaces," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 4, pp. 742–750, Apr. 2007.

[9] G. Bin, X. Gao, Z. Yan, B. Hong, and S. Gao, "An online multi-channel SSVEP-based brain-computer interface using a canonical correlation analysis method," *Journal of Neural Engineering*, vol. 6, no. 4, p. 046002, Jun. 2009.

[10] A. Materka and M. Byczuk, "Alternate half-field stimulation technique for SSVEP-based brain-computer interfaces," *IET Electronic Letters*, vol. 42, no. 6, pp. 321–322, Mar. 2006.

[11] P. Poryzala, M. Byczuk, and A. Materka, "Modular, plugin-based computer software for on-line analysis of electroencephalographic data in Brain-Computer Interfaces," in *II Forum Innowacji Mlodych Badaczy*, Nov. 2011.

[12] M. Byczuk, P. Poryzala, and A. Materka, "On possibility of stimulus parameter selection for SSVEP-based brain-computer interface," in *Man-Machine Interactions 2 (Advances in Intelligent and Soft Computing)*, ser. Advances in Intelligent and Soft Computing, T. Czachorski, S. Kozielski, and U. Stanczyk, Eds. Springer, 2011, vol. 103, pp. 57–64.