Audio and Communications Signal Processing Group

HEAD OF THE GROUP RESEARCH REPORT

The Audio and Communications Signal Processing Group (known by their acronym GTAC from its Spanish name Grupo de Tratamiento de señal en Audio y Comunicaciones) has developed its research during the scholar year 2020-21 mainly on active noise control, spatial audio perception and rendering, and sound quality improvement for multi-channel audio systems. GTAC has carried out several research projects and has published their most relevant results in relevant scientific journals and conference proceedings. In particular, the national projects "Dynamic Acoustic Networks for Changing Environments (DANCE)" and "Intelligent Spatial Audio Synthesis and Customization (ISLA-THESON)", which are in halfway through their completion,and the regional project "Smart Social Computing and Communication (COMTACTS)".

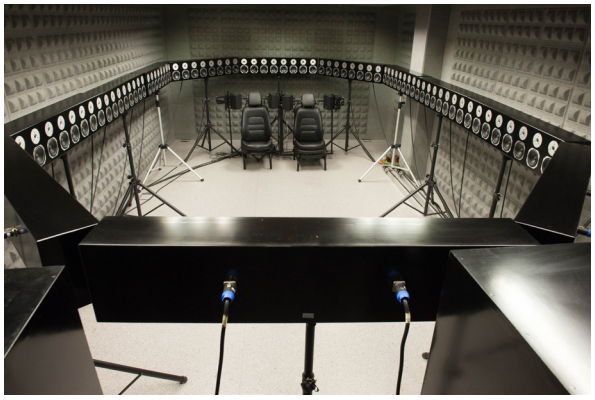


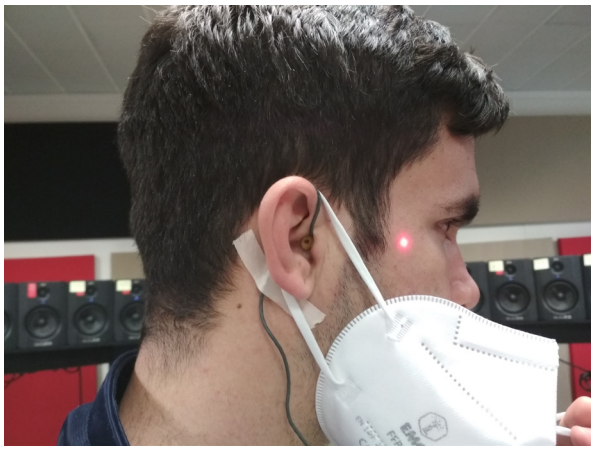
Figure 1. Listening room overview.

On the other hand, the Cátedra Telefonica-UPV project "Sound-Aided Smart Environments for the City, Home and Nature (SSEnCe)" has ended with great success, achieving their objectives by creating a demonstrator that allows detecting and classifying acoustic events, for home environments, Smart Cities and Natural parks. More details of the projects' achievements are shown at the "Ongoing Projects" section.

With regards to the GTAC audio facilities, it comprises two main audio laboratories. A large listening room of 40 m^2, totally equipped with audiovisual and control instrumentation (see Fig.1). This laboratory includes 96 loudspeakers to render an acoustic field over this large listening area. Moreover, car seats are placed in this room to create local quiet zones in enclosures (such as a cabin of a public transport) over a distributed network composed of acoustic nodes.

On the other hand, the laboratory for perceptual spatial sound of Fig. 2 allows measuring Head-Related Transfer Functions (HRTF) of any person with very high precision, in such a way that spatial sound can be rendered to that particular person with high fidelity. The HRTF is somehow a personal acoustic fingerprint that changes from one person to another. By using individualized HRTFs, we can generate a virtual sound that is indistinguishable from reality. As it can be seen from Fig.2 a), the loudspeaker array is formed by a 4-meter-diameter circular array of 72 loudspeakers placed in the same horizontal plane, plus two sets of 8 loudspeakers, one placed in the ceiling and one on the floor. Fig. 2 b) shows a listener with miniature microphones inserted in the ears.

(a)



(b)

Figure 2. Acoustic measurements of the HRTF with miniature microphones inserted in the ears: a) General view of the perceptual spatial sound laboratory when an HRT is measured; b) Miniature microphone detail.

1.- Project activities

In the following we describe the main ongoing projects that are being developed by GTAC researchers.

***Title*: DYNAMIC ACOUSTIC NETWORKS FOR CHANGING ENVIRONMENTS (DANCE)**

Webpage: www.dance.upv.es

Funded by: Spanish Ministry of Science, Innovation and University. 2019-2022.

DANCE is a coordinated project that will develop distributed algorithms and systems to deal with different audio applications under the common frame of dynamic scenarios. Some of their tasks are: self-localization of nodes' positions, estimation of dynamic room impulse responses (RIRs) and inverse filters, fast adaptation and/or implementation over a distributed and heterogeneous network, characterization and control strategies adapted to the environments where control or listening points may vary with time, development of multiuser perceptual equalization methods to improve the listening experience in presence of undesired ambient noises. Additionally, emerging computing tools are used to meet the real-time requirements of audio rendering and control in time-varying scenarios.

The DANCE project includes the development of two testbeds in the GTAC audio laboratory. The first one employs sub-band filtering and optimized filter bank computation in the time domain for the design of personal sound zones (PSZ). The aim is to render a target soundfield in the "bright" zone while having control over the mean acoustic energy in the "dark" (quiet) zone (see Fig. 3). The PSZ system has been adapted and optimized according to the frequency content of the audio signals, the characteristics of the room and the typology and location of the transducers used. Examples are: watching TV and simultaneously listening to different languages

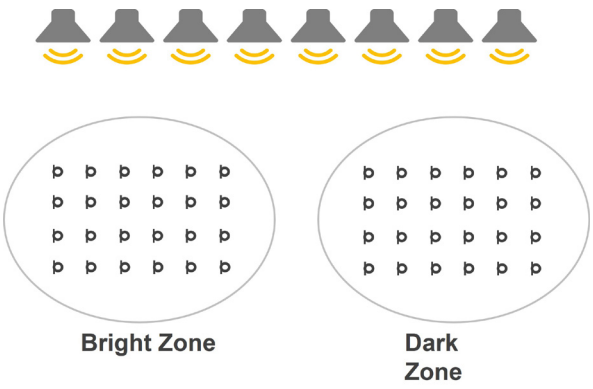


Figure 3. Layout of the PSZ system. Loudspeakers and microphone setup for the bright and dark zones.

in different positions, improving the listening experience in any room, tracking the listener over the home.

The second testbed consists in a massive multichannel noise reduction for open-plan offices. The aim is to reduce the annoyance caused by the ambient noise and speech produced by other workers in open working spaces through their masking with pleasant sounds. Masking implies directing a pleasant sound (such as waterfall, birdsongs, non-voiced music) to a certain zone, such that the pleasant sound makes the annoying noise almost inaudible in that area thanks to the masking properties of the human hearing system.

Figure 4. Masking system.

***Title*: INTELLIGENT SPATIAL AUDIO: SYNTHESIS AND CUSTOMIZATION (ISLA-THESON)**

Funded by: Spanish Ministry of Science, Innovation and University. 2019-2022.

The sound industry has been experiencing profound changes in recent years under the perspective of three complementary approaches: the individual, the group and the contents. Due the advances in virtual reality, mobile devices, video games and innmersive 3D movies, the spatial audio is today a discipline that attracts the attention of the industry. In this context, spatial audio systems try to accurately recreate the acoustic sensations that a listener would perceive within a real listening environment. Moreover, the use of headphones has spread enormously, and the need to reproduce highly realistic spatial sound through them is a great opportunity for the industry. For a very immersive experience, the sound must be customized for each individual based on their anatomy, in particular the head and pinna shape, which define their particular Head-Related Transfer Function (HRTF). Measuring a subject's HRTF is still a costly process that requires specialized facilities and finding an indirect way to get individualized HRTF is required. At ITEAM, we have built a new facility to measure HRTFs of real subjects in an efficient way (Fig.2). By employing Deep Learning

techniques and photographs of the ear/head, we have achieved an HRTF personalization of better quality than previous methods. Previously, a new system has been constructed for the capture and extraction of individual anthropometric parameters from photographs. To this end, work is being done on the creation of 3D models through mobile devices that are equipped with depth cameras (see Fig. 5). The results obtained by combining both objective measurements (individual HRTF and anthropometric parameters) with deep learning techniques, can be evaluated by means of subjective perceptual tests. By using an individualized HRTF, we can generate a virtual sound indistinguishable from reality. This will in turn allow mobile devices to incorporate personalized responses for their direct application in 3D sound, virtual and augmented reality, video games, etc.

Figure 5. Creation of a 3D model for HRTF characterization.

On the other hand, the sound and entertainment industry has been redirected during the recent years to big live shows, where the spatialization of sound is still a challenge and an opportunity for using sound field synthesis algorithms to recreate virtual spaces. Array processing techniques should be developed to control the sound in different listening areas while synthesizing the different live sound objects (musicians, actors, presenters, effects, etc.), adapting the synthesis of each object to its own movement and achieving greater realism over the audience. Other scenarios such as museums, exhibitions, restaurants or smart homes would also benefit from the creation of independent audio zones, using similar techniques employing loudspeaker and sensor arrays.

Finally, from the contents point of view, this subproject will work on creating new methods for the analysis of audio and music based on Machine Learning, with application to synchronized audiovisual effects and live enriched events. The aim is to develop Machine Learning algorithms able to extract features from music and enable the synchronization of 3D animations, lights, or lasers with the music.

***Title*: SMART SOCIAL COMPUTING AND COMMUNICATION (in Spanish: COMUNICACIÓN Y COMPUTACIÓN INTELIGENTES Y SOCIALES - COMTACTS)**

Webpage: www.comtacts.upv.es

Funded by: Prometeo Call. Regional Government – Generalitat Valenciana. 2019-2023.

The advances made in the field of distributed computing and the hardware-software available right now make possible to develop powerful systems to process and exchange information, and at the same time, able to interact with the environment through numerous sets of transducers. These transducers, in turn, provide an ever-increasing volume of signals and data, making possible a more precise knowledge of the social and physical environment of the human beings' daily life.

On the other hand, let us consider the boom in applications arising from computing and communication devices for personal use, and their massive use with the advance of communications; some highlighted applications are human-machine interaction, control systems, location and tracking systems, telepresence, automatic classification, high-speed communications, diagnostic assistance systems, etc. Within this framework, intelligent and social computing and communication is defined as the hybrid mix of the two disciplines in order to face challenges of high socio-economic interest. Science is used for the purpose of communications and computing, but taking into account ubiquity, versatility, scalability, efficiency and cooperative processing of heterogeneous computing and data acquisition device networks.

COMTACTS project considers the physical aspects of computing, signal processing, energy consumption, technology, communication, etc., particularly in distributed, collaborative scenarios where massive and heterogeneous data are provided. In this way, COMTACTS addresses the design, development and implementation of products, systems, programs and algorithms for signal processing and communications, which make use of state-of-the-art architectures, advanced computing and efficient communications within the framework of intelligent computing and communication aimed at tackling social challenges.

Title: SOUND-AIDED SMART ENVIRONMENTS FOR THE CITY, HOME AND NATURE (SSEnCe)

Webpage: www.sound-aided-IOT.webs.upv.es

Funding entity and duration: Cátedra Telefónica-UPV. 2017-2020

The SSEnCe project aims to encourage the development and dissemination of real and practical prototypes focused on the concept of intelligence for the Internet of Things (IoT). Particularly, the project has developed applications mainly addressed to obtain acoustic information of the environment. A second main objective of this project has been the creation of an observatory of technological demonstrators developed by national and international research groups related to the acoustic-aided IoT.

We have developed within the frame of the project a demonstrator of an environmental sound classificafier (ESC) of city sounds based on a wireless acoustic sensor network (WASN) whose nodes are low-cost devices and whose scheme is shown in Fig.6. The WASN recognizes a set of sound events or classes from urban environments. Their nodes are formed by Raspberry Pi devices equipped with outdoor microphones, and they not only record the ambient sound, but can also process and recognize a sound event by means of deep learning (convolutional neural network (CNN) model in Fig.3). In our WASN, the nodes send the resulting probability of every sound class to the server, so the data can be displayed in a map. Such WASNs have many advantages as monitoring system: they are cheap compared to other monitoring systems, they can be easily deployed, and they can work day and night. An additional advantage of our WASN is that uses the open standard FIWARE in their communication network, so the whole system can be replicated without the need of proprietary software or hardware. In order to obtain a classification model adjusted to the city of Valencia, a database that collects different clips related to traffic noise has been recorded in different locations of Valencia. This database called VLCSound, collects different audios that have been previously validated and labeled within the following classes: Traffic, Siren, Horn and Noisy. The goal is to continue increasing VLCSound database to have a robust classification model.

2.- Research results

The most important results of the GTAC publications over the past year are summarized in the following. For a more detailed description, visit our webpage: www.gtac.upv.es where a complete list of projects and papers can be found.

2.1.- Featured Journal publications

Affine Projection Algorithm Over Acoustic Sensor Networks for Active Noise Control. *Miguel Ferrer, Maria de Diego, Gema Piñero, Alberto Gonzalez, IEEE/ACM Trans. on Audio, Speech and Language Processing, vol. 29, pp. 448 - 461, 2020. DOI: 10.1109/TASLP.2020.3042590*

Abstract: Acoustic sensor networks (ASNs) are an effective solution to implement active noise control (ANC) systems by using distributed adaptive algorithms. On one hand, ASNs provide scalable systems where the signal processing load is distributed among the network nodes. On the other hand, their noise reduction performance is comparable to that of their respective centralized processing systems. In this sense, the distributed multiple error filtered-x least mean squares (DMEFxLMS) adaptive algorithm has shown to obtain the same performance than its centralized counterpart as long as there are no

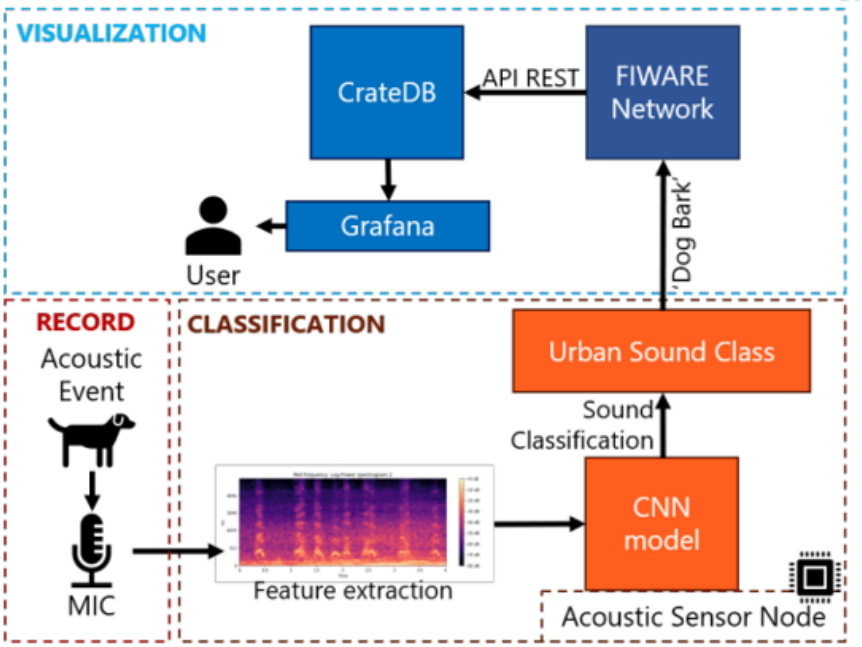


Figure 6. Scheme of the WASN used for cities' sound classification.

communications constraints in the underlying ASN. Regarding affine projection (AP) adaptive algorithms, some distributed approaches that are approximated versions of the multichannel filtered-x affine projection (MFxAP) algorithm have been previously proposed. These AP algorithms can efficiently share the processing load among the nodes, but at the expense of worsening their convergence properties. In this paper we develop the exact distributed multichannel filtered-x AP (EFxAP) algorithm, which obtains the same solution as that of the MFxAP algorithm as long as there are no communications constraints in the underlying ASN. In the EFxAP algorithm each node can compute a part or the entire inverse matrix needed by the centralized MFxAP algorithm. Thus, we propose three different strategies that obtain significant computational saving: 1) Gauss Elimination, 2) block LU factorization, and 3) matrix inversion lemma. As a result, each node computes only between 25% - 60% of the number of multiplications required by the direct inversion of the matrix. Regarding the performance in transient and steady states, the EFxAP exhibits the fastest convergence and the highest noise level reduction for any size of the acoustic network and any projection order of the AP algorithm compared to the DMEFxLMS and two previously reported distributed AP algorithms.

Video-Based System for Automatic Measurement of Barbell Velocity in Back Squat. *Basilio Pueo; José Javier López Monfort; José Manuel Mossi García; Adrián Colomer; Jose M. Jimenez-Olmedo. Sensors 2021, vol. 21 (3), 925. DOI: 10.3390/s21030925.*

Abstract: Velocity-based training is a contemporary method used by sports coaches to prescribe the optimal loading based on the velocity of movement of a load lifted. The most employed and accurate instruments to monitor velocity are linear position transducers. Alternatively, smartphone apps compute mean velocity after each execution by manual on-screen digitizing, introducing human error. In this paper, a video-based instrument delivering unattended, real-time measures of barbell velocity with a smartphone high-speed camera has been developed. A custom image-processing algorithm allows for the detection of reference points of a multipower machine to autocalibrate and automatically track barbell markers to give real-time kinematic-derived parameters. Validity and reliability were studied by comparing the simultaneous measurement of 160 repetitions of back squat lifts executed by 20 athletes with the proposed instrument and a validated linear position transducer, used as a criterion. The video system produced practically identical range, velocity, force, and power outcomes to the criterion with low and proportional systematic bias and random errors. Our results suggest that the developed video system is a valid, reliable, and trustworthy instrument for measuring velocity and derived variables accurately with practical implications for use by coaches and practitioners.

On the performance of a GPU-based SoC in a distributed spatial audio system. *Jose A. Belloch, José M. Badía, Diego F. Larios, Enrique Personal, Miguel Ferrer, Laura Fuster, Mihaita Lupoiu, Alberto Gonzalez, Carlos León, Antonio M. Vidal, Enrique S. Quintana-Ortí. The Journal of Supercomputing, vol. 77, 6920–6935, 2021. DOI: 10.1007/s11227-020-03577-4.*

Abstract: Many current system-on-chip (SoC) devices are composed of low-power multicore processors combined with a small graphics accelerator (or GPU) offering a trade-off between computational capacity and low-power consumption. In this context, spatial audio methods such as wave field synthesis (WFS) can benefit from a distributed system composed of several SoCs that collaborate to tackle the high computational cost of rendering virtual sound sources. This paper aims at evaluating important aspects dealing with a distributed WFS implementation that runs over a network of Jetson Nano boards composed of embedded GPU-based SoCs: computational performance, energy efficiency, and synchronization issues. Our results show that the maximum efficiency is obtained when the WFS system operates the GPU frequency at 691.2 MHz, achieving 11 sources-per-Watt. Synchronization experiments using the NTP protocol show that the maximum initial delay of 10 ms between nodes does not prevent us from achieving high spatial sound quality.

FIWARE based low-cost wireless acoustic sensor network for monitoring and classification of urban soundscape. *Pau Arce, David Salvo-Gutiérrez, Gema Piñero, Alberto Gonzalez, Computer Networks, vol. 9 (196), 2021. DOI: 10.1016/j.comnet.2021.108199.*

Abstract: This work presents a wireless acoustic sensor network (WASN) that monitors urban environments by recognizing a given set of sound events or classes. The nodes of the WASN are Raspberry Pi devices that not only record the ambient sound, but also detect and recognize different sound events. All the signal processing tasks, from the recording to the classification carried out by a convolutional neural network (CNN), are run on Raspberry Pi devices. Due to the low cost of the proposed acoustic nodes, the system exhibits a very high potential scalability. Regarding the underlying WASN, it has been designed according to the open standard FIWARE, thus the whole system can be deployed

without the need of proprietary software. Regarding the performance of the sound classifier, the proposed WASN achieves similar accuracy compared to other WASNs that make use of cloud computing. However, the proposed WASN significantly minimizes the network traffic since it does not exchange audio signals, but only contextual information in form of labels. On the other hand, most of the time the class reported by the WASN nodes is the "background" soundscape, which usually contains no event of interest. This is the case when monitoring the soundscape of big avenues, where four events have been identified: "traffic", "siren", "horn" and "noisy vehicles", being the "traffic" class associated to the background soundscape. In this paper, the use of a simple pre-detection stage prior to the CNN classification is proposed, with the aim of saving computation and power consumption at the nodes. The pre-detection stage is able to differentiate the other three relevant sounds from the "traffic" and activates the classifier only when some of these three events is likely occurring. The proposed pre-detection stage has been validated through data recorded in the city of Valencia (Spain), achieving a reduction of the Raspberry Pi CPU's usage by a factor of six.

Maximum likelihood low-complexity GSM detection for large MIMO systems. *Victor M.Garcia-Molla, F.J.Martínez-Zaldívar, M.Angeles Simarro, Alberto Gonzalez, Signal Processing, vol. 175, 2020. DOI: 10.1016/j.sigpro.2020.107661.*

Abstract: Hard-Output Maximum Likelihood (ML) detection for Generalized Spatial Modulation (GSM) systems involves obtaining the ML solution of a number of different MIMO subproblems, with as many possible antenna configurations as subproblems. Obtaining the ML solution of all of the subproblems has a large computational complexity, especially for large GSM MIMO systems. In this paper, we present two techniques for reducing the computational complexity of GSM ML detection. The first technique is based on computing a box optimization bound for each subproblem. This, together with sequential processing of the subproblems, allows fast discarding of many of these subproblems. The second technique is to use a Sphere Detector that is based on box optimization for the solution of the subproblems. This Sphere Detector reduces the number of partial solutions explored in each subproblem. The experiments show that these techniques are very effective in reducing the computational complexity in large MIMO setups.

An Efficient Implementation of Parallel Parametric HRTF Models for Binaural Sound Synthesis in Mobile Multimedia. *Jose A. Belloch, German Ramos, Jose M. Badia, Maximo Cobos, IEEE Access, vol. 8, 49562 - 49573, 2020. DOI: 10.1109/ACCESS.2020.2979489.*

Abstract: The extended use of mobile multimedia devices in applications like gaming, 3D video and audio reproduction, immersive teleconferencing, or virtual and augmented reality, is demanding efficient algorithms and methodologies. All these applications require real-time spatial audio engines with the capability of dealing with intensive signal processing operations while facing a number of constraints related to computational cost, latency and energy consumption. Most mobile multimedia devices include a Graphics Processing Unit (GPU) that is primarily used to accelerate video processing tasks, providing high computational capabilities due to its inherent parallel architecture. This paper describes a scalable parallel implementation of a real-time binaural audio engine for GPU-equipped mobile devices. The engine is based on a set of head-related transfer functions (HRTFs) modelled with a parametric parallel structure, allowing efficient synthesis and interpolation while reducing the size required for HRTF data storage. Several strategies to optimize the GPU implementation are evaluated over a well-known kind of processor present in a wide range of mobile devices. In this context, we analyze both the energy consumption and real-time capabilities of the system by exploring different GPU and CPU configuration alternatives. Moreover, the implementation has been conducted using the OpenCL framework, guarantying the portability of the code.

2.2.- Featured Conference Proceedings

- ***Perceptual Active Equalization of Multi-frequency Noise.*** *Juan Estreder, Gema Piñero, Miguel Ferrer, Maria de Diego, Alberto Gonzalez 18th International Conference on Signal Processing and Multimedia Applications (SIGMAP), Online, 2021.*

- ***Subjective analysis of speech privacy using speech masking in open-plan offices.*** *Laura Fuster, Maria de Diego, Gema Piñero, Alberto Gonzalez, Miguel Ferrer 27th International Congress on Sound and Vibration (ICSV27) , Online, 2021.*

- ***Low Complexity Near-ML Sphere Decoding based on a MMSE ordering for Generalized Spatial Modulation.*** *M. Angeles Simarro, Víctor M. García, Francisco J. Martínez-Zaldívar, Alberto Gonzalez, 31th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2020).*