Expert Systems with Applications 42 (2015) 5607–5620

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

On the performance of multi-GPU-based expert systems for acoustic localization involving massive microphone arrays

Jose A. Belloch[a,*], Alberto Gonzalez[a], Antonio M. Vidal[b], Maximo Cobos[c]

[a] *Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain*
[b] *DSIC Department, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain*
[c] *Computer Science Department, Universitat de València, Av. de la Universitat s/n, 46100 Burjassot, Valencia, Spain*

ARTICLE INFO

Article history:
Available online 12 March 2015

Keywords:
Sound source localization
Steered Response Power
Microphone arrays
Graphics Processing Units

ABSTRACT

Sound source localization is an important topic in expert systems involving microphone arrays, such as automatic camera steering systems, human–machine interaction, video gaming or audio surveillance. The Steered Response Power with Phase Transform (SRP-PHAT) algorithm is a well-known approach for sound source localization due to its robust performance in noisy and reverberant environments. This algorithm analyzes the sound power captured by an acoustic beamformer on a defined spatial grid, estimating the source location as the point that maximizes the output power. Since localization accuracy can be improved by using high-resolution spatial grids and a high number of microphones, accurate acoustic localization systems require high computational power. Graphics Processing Units (GPUs) are highly parallel programmable co-processors that provide massive computation when the needed operations are properly parallelized. Emerging GPUs offer multiple parallelism levels; however, properly managing their computational resources becomes a very challenging task. In fact, management issues become even more difficult when multiple GPUs are involved, adding one more level of parallelism. In this paper, the performance of an acoustic source localization system using distributed microphones is analyzed over a massive multichannel processing framework in a multi-GPU system. The paper evaluates and points out the influence that the number of microphones and the available computational resources have in the overall system performance. Several acoustic environments are considered to show the impact that noise and reverberation have in the localization accuracy and how the use of massive microphone systems combined with parallelized GPU algorithms can help to mitigate substantially adverse acoustic effects. In this context, the proposed implementation is able to work in real time with high-resolution spatial grids and using up to 48 microphones. These results confirm the advantages of suitable GPU architectures in the development of real-time massive acoustic signal processing systems.

1. Introduction

Microphone arrays are commonly employed in many signal processing tasks, such as speech enhancement, acoustic echo cancellation or sound source separation (Brandstein & Ward, 2001). The localization of broadband sound sources under high noise and reverberation is another challenging task in multichannel signal processing, being an active research topic with applications in human–computer interfaces (Kodagoda & Sehestedt, 2014), teleconferencing (Wang, Wang, & Kim, 2011) or emergency units (Calderoni, Ferrara, Franco, & Maio, 2015). Microphone arrays may follow a given geometry, such as spherical arrays (Huang & Wang, 2014), or may be distributed. Algorithms for sound source localization can be broadly divided into indirect and direct approaches (Madhu & Martin, 2008). Indirect approaches usually follow a two-step procedure: they first estimate the *Time Difference Of Arrival* (TDOA) (Chen, Benesty, & Huang, 2006) between microphone pairs, and, afterwards, they estimate the source position based on the geometry of the array and the estimated delays. On the other hand, direct approaches perform TDOA estimation and source localization in one single step by scanning a set of candidate source locations and selecting the most likely position as an estimate of the real source location. Although the computation of TDOAs usually requires time synchronization, new approaches are being developed to avoid this limitation (Xu, Sun, Yu, & Yang, 2013). Most localization algorithms are based on the *Generalized Cross-Correlation* (GCC) (Knapp & Carter, 1976), which is calculated by using the inverse Fourier transform of the

* Corresponding author. Tel.: +34 655436190.
E-mail addresses: jobelrod@iteam.upv.es (J.A. Belloch), agonzal@dcom.upv.es (A. Gonzalez), avidal@dsic.upv.es (A.M. Vidal), maximo.cobos@uv.es (M. Cobos).

http://dx.doi.org/10.1016/j.eswa.2015.02.056
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

weighted cross-power spectral density of the signals. The *Steered Response Power-Phase Transform* (SRP-PHAT) algorithm is a direct approach that has been shown to be very robust in adverse acoustic environments (DiBiase, Silverman, & Brandstein, 2001, chap. 8). The algorithm is usually interpreted as a beamforming-based approach that searches for the candidate position that maximizes the output of a steered delay-and-sum beamformer.

The CUDA platform (CUDA, 2015) provides a computing framework that enables the use of Graphics Processing Units (GPUs) in applications beyond image processing (Liu, Schmidt, Voss, & Muller-Wittig, 2007; Zhao & Lau, 2013). GPUs are high parallel programmable co-processors that provide efficient computation when the needed operations are properly parallelized. Programming a GPU efficiently requires having good knowledge of both the underlying architecture and the mechanisms used by GPUs to distribute their tasks among their processing units. Since the appearance of CUDA programming, many researchers in different areas have made use of it to achieve better performances in their respective fields. For example, well-known computational cores have also been adapted to a GPU computing framework, such as LU factorization (Dazevedo & Hill, 2012), matrix multiplication (Matsumoto, Nakasato, Sakai, Yahagi, & Sedukhin, 2011) or the Boltzmann equation (Kloss, Shuvalov, & Tcheremissine, 2010). In audio and acoustics, several works demonstrate the potential of GPUs for carrying out audio processing tasks. For example, the implementation of a multichannel room impulse response reshaping algorithm was carried out in Mazur, Jungmann, and Mertins (2011), and implementations of adaptive filtering algorithms were presented in Schneider, Schuh, and Kellermann (2012), Lorente et al. (2012), Lorente, Ferrer, De Diego, Belloch, and Gonzalez (2013) and Lorente, Ferrer, De Diego, and Gonzalez (2014). GPU-based room acoustics simulation was carried out in Savioja (2010), Southern, Murphy, Campos, and Dias (2010), Webb and Bilbao (2011) and Hamilton and Webb (2013). One of the main contributions within this field was carried out in Savioja, Välimäki, and Smith (2011), where improved performances in additive synthesis, Fourier transform and convolution in the frequency domain were presented. A comparison between CPU and GPU performance for a simple crosstalk canceller is presented in Belloch, Gonzalez, Martinez-Zaldivar, and Vidal (2011). Similarly, a binaural audio application with massive audio processing that was fully implemented on a GPU is presented in Belloch, Ferrer, Gonzalez, Martinez-Zaldivar, and Vidal (2013a). GPUs are also used in Vanek, Trmal, Psutka, and Psutka (2012) and in Bradford, Ffitch, and Dobson (2011) for evaluating the likelihood function in automatic speech recognizers and for sliding phase vocoder, respectively.

The use of GPUs for implementing sound source localization algorithms has also recently been tackled in the literature. The time performances of different localization algorithms implemented on GPU were reported in Peruffo Minotto, Rosito Jung, Gonzaga da Silveira, and Lee (2012) and Liang et al. (2012). In fact, although different implementations of the SRP-PHAT in the time-domain and frequency-domain are analyzed in Peruffo Minotto et al. (2012), their results mainly focus on pure computational issues and do not discuss how localization performance is affected by using different numbers of microphones or a finer spatial grid. In Seewald, Gonzaga, Veronez, Minotto, and Jung (2014), the SRP-PHAT algorithm is implemented over two Kinects for performing sound source localization. In the same work, the algorithm only estimates the relative source direction instead of providing the absolute source position and the implementation is evaluated on different GPUs that belong to the old-fashioned Fermi (CUDA, 2015).

One of our previous works (Belloch, Gonzalez, Vidal, & Cobos, 2013b) analyzed the performance of a 2-D SRP-PHAT implementation with different Nvidia GPU architectures. The present paper extends that work in various aspects. First, 3-D source localization is considered, leading to a significant increase in the required computational cost. Second, the system considered in this work makes use of multiple GPUs, facing new challenges in parallelization and resource management. Finally, this paper provides a deeper analysis of the influence of the acoustic environment and the number of microphones in the final performance. As a result, this paper is aimed at demonstrating how localization systems using a high number of microphones distributed within a room can perform sound source localization in real time under adverse acoustic environments by using GPU massive computation resources. Specifically, the well-known SRP-PHAT algorithm is considered here. Note that coarse-to-fine search strategies have been proposed to overcome many of the processing limitations of SRP-PHAT (Do & Silverman, 2007; Marti, Cobos, & Lopez, 2013; Said, Lee, & Kalker, 2013). However, while these strategies provide more efficient ways to explore the localization search volume, they only provide better performance than the conventional SRP-PHAT when the number of operations is restricted. Thus, the performance of the conventional SRP-PHAT with fine spatial grids is usually considered as an upper bound in these cases.

Relevant parameters that affect the computational cost of the algorithm (number of microphones and spatial resolution) are analyzed, showing their influence on the localization accuracy in different situations. We also discuss the scalability of the algorithm when multi-GPU parallelization issues are considered. This paper highlights the need for massive computation in order to achieve high-accuracy localization in adverse acoustic environments, taking advantage of GPUs to fulfill the computational demand of the system.

In comparison with the implementation presented in Seewald et al. (2014), we design our application to achieve maximum performance on GPUs making use of the Kepler architecture GK110 (K20, 2014) (See Appendix A for details). This architecture can be found on the Tegra K1 (TK1) systems-on-chip (SoC), embedded in the Jetson development kit (DevKit) (Jetson, 2015), and it is becoming widespread in current mobile devices such as Google's Nexus 9 tablet (Nexus, 2015). Thus, the proposed implementation can be successfully adapted to work properly on GPUs that are currently embedded in mobile devices.

The paper is structured as follows. Section 2 briefly describes the basic SRP-PHAT localization algorithm that will be used throughout this paper. Section 3 presents the implementation of the algorithm on multi-GPU systems. The proposed acoustic environments for real-time sound source localization are presented in Section 4, describing the experiments conducted for studying the performance of the method in a real application context. The computational performance of the different multi-GPU implementations are also analyzed. Finally, Section 5 provides some concluding remarks. Two Appendixes are provided in order to facilitate the understanding of the parallelization techniques that are used throughout this article.

2. Sound source localization: SRP-PHAT algorithm

Consider the output from microphone $l, m_l(t)$, in an M microphone system. The Steered Response Power (SRP) at the spatial point $\mathbf{x} = [x, y, z]^T$ for a time frame n of length T_L can then be defined as

$$P_n(\mathbf{x}) \equiv \int_{nT_L}^{(n+1)T_L} \left| \sum_{l=1}^{M} w_l m_l(t - \tau(\mathbf{x}, l)) \right|^2 dt, \qquad (1)$$

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase (DiBiase, 2000) showed that the SRP