Spatial Audio Coding and Machine Learning

S

Karim Dabbabi

Faculty of Sciences of Tunis, Research Unit of Analysis and Processing of Electrical and Energetic Systems, University of Tunis El Manar, Tunis, Tunisia

INTRODUCTION

The development and improvement of spatial sound rendering techniques is the result of the advent of consumer mixed reality products, which are becoming addressed to a large user base. In this regard, acoustics processing and modeling in the field of audio analysis has found rapid inception and adoption using many machine and deep learning methods to meet the domain-specific requirements of acoustic research. This requires new approaches and architectures to be adapted. Such methods were inspired from image processing and adapted to audio processing, such as adversarial approaches and 2D convolutional operators.

In fact, the challenges posed by the processing of acoustic signals are numerous which are linked on the one hand to their nature and their representations as well as to the nature of the anisotropy of their time-frequency representations with the short-term Fourier transform, and on the other hand to the multi-scale nature of musical events as well as to the effect of psychoacoustics. Spatial audio also has its specific complexity that influences detection performance for humans and machines (Zieliński, Lee, Antoniuk, & Dadan, 2020). It is mainly devoted to the localization of sound ensemble around, in front of or behind a listener. On the Internet, human objects took on the role of remote testing, but their accuracy was low (Gabrielli, Fazekas, & Nam, 2021). In contrast, very high classification results have been obtained in tests with deep learning methods under unknown conditions (Gabrielli, Fazekas, & Nam, 2021).

In this chapter, spatial audio coding standards will be discussed with the aim of presenting the importance of their applicability and showing their strengths and weaknesses Among these standards, a focus will be made on spatial audio coding techniques (SAC), followed by a brief passage on the psychoacoustic principle of spatial sound, then the historical reception of multichannel audio will be presented by listing the main approaches. After that, the suggested SAC techniques will exposed. The way to improve the quality accuracy for the reconstructed audio, and the evaluation of the quality of the reconstructed audio signal will be given successively at the end of this first section. MPEG standards for encoding multi-channel audio signal, such as MPEG Surround, MPEG Spatial Audio Object Encoding, and MPEG-H 3D Audio Encoding, will be introduced in the second section. The applications of machine learning (ML) in acoustics and their exploration for spatial sound scenes will then be successively integrated and analyzed. At the end, other research directions in spatial audio and machine learning (ML) are suggested and analyzed.

DOI: 10.4018/978-1-7998-9220-5.ch149

BACKGROUND

Nowadays, many technological inventions have been made and integrated into the market, such as three-dimensional (3D) audio technology, also called spatial audio (Rumsey, F., 2001). The latter has many application areas, such as digital audio entertainment media like ultrahigh definition television (UHDTV), many other generations of television broadcasting, etc. As for the UHDTV standard, up to 20 multiple speakers have been explored to provide realistic 3D audio perception to users. Thus, a single audio channel feeds each speaker; therefore, multi-channel audio signals will be requested. Some broadcasting companies which have continuously adopted these audio chain technologies in recording, transmission and production include the BBC, UK and NHK, Japan.

In audio spatial, the active component that plays a major role is perceptual audio coding (Pan, D., 1995; Painter, T., & Spanias, A., 2000; Bosi, M., & Goldberg, R. E., 2002; Brandenburg, Faller, Herre, Johnston, & Kleijn, 2014).

The latter has been developed in such a way that it can compress the size of audio data incredibly so that the properties of the audio signal that would not be detected by our hearing system simply have to be removed. This task is carried out on the basis of knowledge about psychoacoustics. Indeed, the emergence of perceptual audio coding refers to the 1990s, when the first and well-known MPEG-1 layer 3 digital audio compression standard, known as MP3, was explored. In fact, spatial audio coding (SAC) (Herre, Faller, Disch, Ertel, Hilpert, Hoelzer, Linzmeier, Spenger, & Kroon,2004; Herre, J., 2004). is one of many other audio coding techniques that have been invented and standardized to accurately model multi-channel audio signals. The goal behind the use of spatial audio coding is not only limited to increasing the compression rate over advanced multi-channel audio coding methods but is being extended to provide more opportunities to be applied in other systems, such as the legacy broadcasting system. Surround MPEG is considered a well-used SAC standard, which has aroused the interest of researchers due to its rich functionalities, such as artistic stereo mixing and binaural rendering. In addition, an option was provided for users to interact and update the audio scene composition and spatial characteristics of the rendered surround sound by integrating object-based audio.

This makes it more interesting audio rendering system. There are many applications that are keen to apply technology based on objects, such as musical reconstruction, games, teleconferencing, sports broadcasting, karaoke system and improving dialogue (Oldfield, Shirley, & Spille, 2014; Oldfield, Shirley & Satongar, 2015; Jot, Smith & Thompson, 2015; Bleidt, Borsum, Fuchs, & Weiss, 2015).

FOCUS OF THE ARTICLE

Standards of Spatial Audio Coding

Spatial Audio Coding (SAC) is more its primary consideration as not being a pure compression method, it is counted as an approach to represent multi-channel audio signals by a lower number of channels (i.e., a (mono) or two (stereo)) with preservation of the spatial properties of the audio signals. Typically, the process responsible for reducing the number of channels is called down-mixing, as shown in Figure 1. For transmission and storage tasks, the encoding of the down-mix signals should be done using compression techniques, such as Universal Speech and Audio Coding (USAC) (Quackenbush, S., & Lefebvre, R.,,2011; ISO/IEC,2012;Neuendorf, Multrus, Rettelbach, Fuchs, Robilliard, Lecomte, & Grill,2013;Oh, E., & Kim, M., 2011)., MPEG 1 layer 3, and Advanced Audio Coding (AAC) (Bosi, Brandenburg, Quackenbush,