Over the last couple of years, Artificial Intelligence (AI) has revolutionized the technology industry. We notice that in everyday life, as AI slowly starts to help us with a number of daily tasks. For instance, AI guards our privacy, making sure nobody unauthorized can unlock our devices while enabling us to easily access them by recognizing our face, fingerprint, or voice. Most commonly, when we say AI, we mean Deep Neural Networks (DNN) which can learn to perform extremely complex tasks by mimicking the structure of the human brain. In a nutshell, the classical way to train the DNN involves showing it millions of question-answer examples, referred to as data and labels. By feeding the questions to the DNN and comparing its answers with the ground truth, we can update the DNN so that it incrementally becomes better at solving a problem. This approach is called supervised learning and it is very effective in various isolated tasks, for which a large training dataset with appropriate labels is available.

Unfortunately, it is not straightforward to apply supervised learning to any given problem. In real-life applications, we often do not have access to reference content in general, which prevents us from using the supervised learning framework. Partially, this is because the labels have to be prepared with a specific purpose in mind, which may just not fully meet our needs for a differently defined task. Secondly, since labeling commonly has to be done manually, it is often not feasible in practice. Another issue concerns performance degradation with out-of-domain data, which happens when the training data is substantially different from the target domain. That can be the case when DNN is trained to understand speech using recordings made with a professional microphone in studio conditions, and then it processes audio from a smartphone microphone on a busy street.

A great deal of these problems can be addressed by adopting a self-supervised learning (SSL) paradigm, in which, rather than trying to capture the question-answer relation, we orient the model towards the understanding of the underlying structure of the data itself. Besides solving the label-related problems, self-supervised learning provides a more meaningful model of the data, which improves the DNN performance in a number of applications. In general, it can be achieved by involving a prediction step in the teaching process. For example, in Natural Language Processing (NLP), during training only a part of a sentence can be presented as input to the DNN, whose goal is then to predict the missing words based solely on sentence context. While SSL has been shown to be highly successful in NLP, it remains underdeveloped in acoustic applications.

Encouraged by the immense success seen in NLP, in the Acoustic Intelligence project, we aim at unlocking the full potential of self-supervised learning in the context of acoustic applications. We intend to address this frontier in several ways. One of them includes innovative universal audio representation, where a DNN model will be pre-trained without specifying its end-purpose. By applying a novel multitask training scheme, we expect to orient the model towards a general understanding of an auditory scene, much like a human does. We will also introduce a novel approach, whose key idea is to exploit the sound events appearing in a recording to help classify the type of an acoustic scene. For example, if we hear birds chirping, barking dogs and children's laughter, we can guess that the scene takes place in a park. The rest of our focus will be oriented towards universal audio signal enhancement. As the bottom line, we will contribute to the creation of intelligent machines, capable of autonomous learning from acoustic signals, such as self-adaptation for performance upholding, even in previously unseen scenarios.

Ultimately, our research can help bring AI from lab conditions to the benefit of society. By lifting some of the limitations of AI, its increased capabilities could e.g. help people to reconnect by enhancing the hearing support technology. The autonomous Acoustic Intelligence has the potential to be applied e.g. for intelligent acoustic assistance, such as providing guidance in the city and warnings about possibly dangerous acoustic events in our surroundings e.g. a passing ambulance. The robot helper aware of its surroundings could react accordingly, e.g. call for help when hearing a body falling on the floor of a solitary house, followed by a moan, but not when the fall occurs on stage and laughter follows the scene. The powerful scanning capabilities of an intelligent acoustic model could also be used to tag the huge audio-video content available on social media. Finally, the Acoustic Intelligence would also find wide application in acoustic signal enhancement, e.g. to denoise low frequency content of speech signal, when a windy outdoor is recognized. The autonomous, self-adapting, intelligent audio signal enhancement would impact acoustic signal processing technologies in numerous ways, such as improving the safety of acoustic biometric security systems or by limiting the influence of the recording device or acoustic surroundings on the performance of downstream tasks.