Popular summary

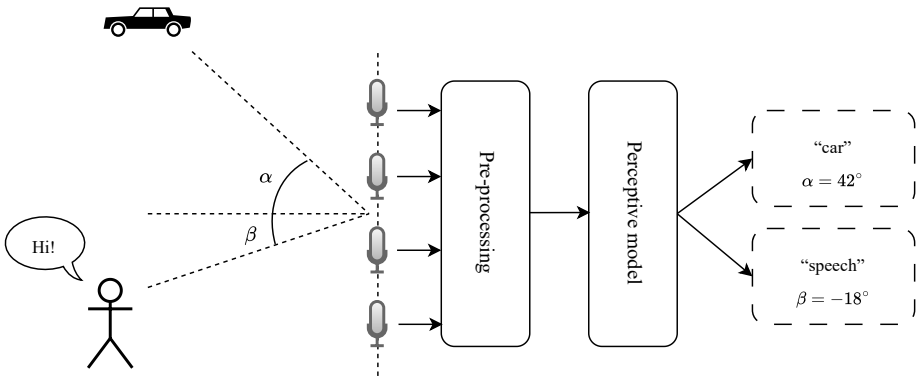
Perception is an important building block in artificial intelligence (AI) systems which allows them to process sensory information from the outside world. The "senses" of an AI system are the sensors that convert physical phenomena into digital signals, e.g. cameras, microphones and antennas. The perception consists of interpreting these signals by extracting the relevant information, such as the semantic content of the images captured by the camera or the words in a speech recording captured by the microphones. Depending on the use case of the system, the information can be used in different ways. For example, an autonomous robot might use perception in order to navigate through its surroundings. In other use cases, perceptive AI systems can assist human decision making, such as in advanced driver-assistance systems used in many modern vehicles.

Today, most state of the art perceptive methods are built using machine learning, i.e. statistical algorithms, or "models", that learn representations of the signals by training on interpreting large amounts of data. In this thesis, a set of new machine learning methods for perception tasks are introduced, with focus on auditory perception. Designing such methods includes training models for classifying different types of sound recordings.

In some scenarios, it is necessary to also be able to localize the sound events in 3D space. Sound source localization has several important applications within the fields of robotics and navigation, and there are also military applications. For this task, it is typically necessary to work with multiple audio recordings from a synchronized microphone array. This is analogous to how humans benefit from having two ears when localizing sounds. Training a localization model for this task therefore involves processing multichannel audio and extracting spatial cues that allows the model to infer the locations of the sound events.

When designing machine learning models for perception, there are several evaluation criteria that ought to be considered. Perhaps most importantly, the inferences made by the system ought to be as accurate as possible. However, just as is the case in human perception, there will always be a small probability of making incorrect inferences. For example, the system might not be able to understand speech with poor pronunciation. Other factors, such as background noise and interference can also make it difficult for the system to understand the data. Therefore it is important to design the training data for the perceptive model in such a way that it contains all the scenarios required to be managed by the system.

When using AI systems in real-world scenarios it is also necessary to consider the limited hardware capabilities of the machine on which the model is being deployed. In recent years, there has been an accelerating trend for models to become more complex and have larger number of trainable parameters. This increases both the memory footprint of the model and possibly also the time required to perform inferences. In this thesis, the perceptive models are therefore analyzed from a computational perspective as well, and solutions for



Conceptual overview of an auditory perceptive system that simultaneously classifies and localizes multiple audio events. Audio signals are recorded by a microphone array and after some pre-processing, they are sent to the perceptive model which infers information about the events and their location. In this thesis, we propose learning-based methods for solving these types of problems.

reducing model complexity are proposed.

In the illustration shown above, we show a simplified example of one of the perceptive tasks considered in the thesis. Here, the goal is to train a model that can detect, classify, localize and track different sound events over time. Most approaches to this problem rely on a pre-processing stage that extracts two types of hand-crafted audio features: 1) spectral features that are good for distinguishing different types of sounds and 2) spatial features that contain information about the location of the sounds source. However, these spatial features do not deal adequately with overlapping sounds and noisy environments, leading to incorrect inferences. To overcome this, a new spatial feature that can be trained to adapt to different environments and learn to separate overlapping sound events is introduced in this thesis. We test this method on real-world audio recordings and the results show that our learning-based approach to feature extraction yields better localization performance than the hand-crafted one.

Machine learning methods have proven to be very successful at solving problems related to perception due to their ability to model reality by learning from data. Nevertheless, the performance of the methods do not only depend on the quality of the data, but also the structure of the models and how the learning procedure is implemented. How should we design the models in order efficiently extract relevant information from the data? One answer is that the model design ought to be adapted to the underlying structures, or symmetries, of the inputs and outputs to the model, which depend not only upon the data, but also on the particular task the model is being trained for. Continue reading the thesis in order to find out how symmetries can be exploited in order to improve the performance of perceptive models.