Reinhold Haeb-Umbach

ISCA Distinguished Lecturer, Term 2021/2022

Professor of Communications Engineering
Paderborn University, Paderborn, Germany

(1) Far-field speech recognition

Recognizing speech uttered at a distance from the microphones, so-called far-field or distant speech recognition, is a major success story of recent years and at present still a very active field of research. It is a key component for the success of digital home assistants. In this talk we open the lid of such devices and have a look at the speech technology inside. In particular, we will discuss the various techniques to achieve reliable speech recognition in the presence of an acoustically very challenging environment. We will discuss modern microphone array processing, dereverberation, source separation, beamforming and multi-condition training techniques, all of them contributing to high-performance far-field ASR. It will become evident that in all these fields, deep learning occupies a critical role, but it will also be seen that a clever combination of signal processing and deep learning leads to highly effective solutions with far less computational demands than pure deep learning solutions.

The lecture is an introduction to the field, touching upon different techniques, but not going into too much detail.

(2) Microphone array signal processing and deep learning for speech enhancement

While multi-channel speech enhancement was traditionally approached by linear or non-linear time-variant filtering techniques, in the last years neural network based solutions have achieved remarkable performance by data-driven learning techniques. Even hybrid methods, which blend traditional signal processing with deep learning, have been proposed in an attempt to combine the best of both worlds: achieving excellent performance while at the same time being resource efficient. In this presentation we discuss signal processing based and neural network based methods to enhance multi-channel speech input, with a focus on automatic speech recognition for applications such as digital home assistants and meeting recognition. This will include, but is not limited to, acoustic beamforming, speech dereverberation and source separation.

This seminar can be easily split in two or three lectures.

(3) Deep generative modeling for disentangling sources of variation in speech

The speech signal is a rich source of information that conveys not only linguistic but also extra/paralinguistic information, as well as information about the acoustic environment. If we were able to disentangle those sources of variation, numerous applications would benefit. For example, robustness to new environments could be greatly improved by segregating the influence of the environment from the data. Voice privacy concerns could be addressed by obfuscating the speaker identity while leaving the linguistic content of the speech untouched. Also, low and zero resource speech technologies would benefit if content induced variations of the speech signal could be separated from speaker induced variations, when learning the acoustic building blocks of speech from the speech signal without supervision.

Today, the most successful techniques for disentangling sources of variation employ deep learning methods, which map the speech signal to embeddings that capture certain sources of variation, while discarding others, using appropriate training objectives. With techniques like contrastive predictive coding, adversarial classifiers or cycle consistency, good disentangling performance can be achieved. The lecture gives an overview of techniques and applications. An emphasis is on combining deep learning with statistical graphical models.

(4) Signal processing and machine learning for acoustic sensor networks

Given the ubiquity of wireless devices, the availability of multiple, spatially distributed microphones is a likely scenario for speech, or more general, acoustic signal processing. Ad hoc array processing allows people to use their own smartphones to form a virtual microphone array or to connect to and enhance a microphone array installed in a room.

Being distributed in an environment an acoustic sensor network promises better signal acquisition than a single compact microphone array, because there is a good chance that a sensor is close to every relevant sound source. The opportunities for signal enhancement are well documented, so are the challenges involved with distributed microphones. Among them are the unknown and possibly time-varying number and spatial arrangement of the microphones, and the lack of a clock synchronization across devices, which breaks with the assumptions made in the overwhelming body of algorithms/literature for multi-channel acoustic signal processing.

This lecture gives an overview of signal processing and machine learning for acoustic sensor networks. We will discuss task distribution, sampling rate synchronization, signal enhancement and acoustic event and scene classification techniques, as well as machine learning techniques to ensure user privacy.