Audio Source Separation using Deep Neural Networks

Maximilian Ammann
University of Augsburg
Augsburg, Germany
maximilian.ammann@student.uni-augsburg.de
Supervisor: Maximilian Schmitt, Shuo Liu

1. EXPOSE

Audio Source Separation deals with the separation of a single audio signal into multiple signals. As early as 1953 the "Cocktail Party Problem" was first mentioned [**?**].

Since then, there have been many approaches to the separation of signals, which have been used in a variety of areas such as automated karaoke, object-specific equalizers, hearing prostheses and robust speech recognition [**?**]. Algorithms can use assumptions such as that the signal is stereo or certain harmonic or redundant compositions exist in the signal to allow separation.

The field of application discussed in this paper is the separation between different voices. Fist a few early approaches will be reviewed to conclude that they are not sufficient enough for most applications.

The results of "Robust principal component analysis" or "Non-negative matrix factorization" will be reviewed shortly to give an overview of the state of the art before the appearance of deep learning.

The next chapter will give an overview of several approaches which have been tried in the past (MLP, CNNs, GANs). This will not contain a detailed description about the network architecture but should give an overview of the research in the past few years. We will also conclude that the most interesting approach is deep clustering. The rough estimations of the goals should be described in this chapter. A major part of this chapter will also be to describe the available data for training and testing the network. Ways to remix the audio should be described in the chapter "Data augmentation".

The main chapter will describe the model I want to use. I want to base my model on the TensorFlow implementation by zhr1201 on Github.com. To be able to describe it in detail the first step is to get his model running and compute the first results. This chapter is not very clear as of now because I do not know how good the results will be and how good the implementation is. It is also not clear whether the implementation matches the paper exactly.

I also want to tweak parameters to achieve better accuracy. The progress of increasing the accuracy should be discussed in the chapter.

In the last part I want to do an evaluation by comparing SDR, IRS, SIR and SAR metrics. I want to test different kinds of voice like female vs. male or child vs. adult. A comparison with one state of the art algorithm should be made if there is data about their accuracy available publicly.

I think implementing the model and gathering the evaluation data will take about 2 month. So I'll have enough time in the last month to write the actual thesis. I will start with the work as soon as the new semester starts as I have to write an other small paper. So I think we can register the thesis in July as Max already mentioned. If I finish earlier than expected I want to evaluate whether the model is usable on embedded hardware. This will be like a outlook for future medial applications.

Seminar: Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany