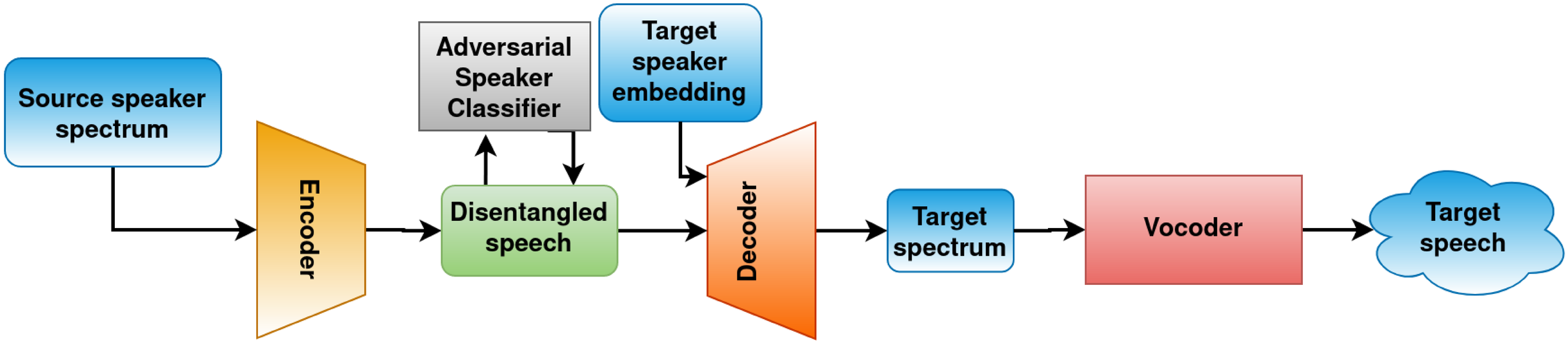
# Non-Parallel Voice Conversion

Student: Jan Brukner*
Supervisor: Jan Černocký
Consultant: Tomi Kinnunen, University of Eastern Finland
*ibrukner@fit.vutbr.cz

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

## Scheme




## Voice Conversion

Voice Conversion (VC) is a task, where the goal is to transform speech of one speaker in such a way, that it sounds as if it was uttered by target speaker. First VC systems were developed using datasets consisting of only parallel utterances with conversion model for each source – target speaker pair. However, parallel datasets are expensive to obtain, therefore non-parallel techniques needed to be developed.
This thesis aims even one step further for so called *one-shot* voice conversion. Not only that one-shot models do not need parallel dataset, but also, once properly trained, they can convert voice from and to any speaker. For this task, state-of-the-art autoencoder based *AutoVC* [1] system was adopted and modified for better results.
Vanilla AutoVC system performs VC by speaker disentanglement (separating speaker dependent and independent information) using properly designed bottleneck and extracts target speaker voice characteristics from speaker embedding – *x-vector*.

## Experiments

Original AutoVC was enhanced with adversarial speaker classifier, that improves level of disentanglement and allows us to use larger bottleneck for better quality of resulting speech.
Usually, VC systems are trained and evaluated only on speech samples recorded in controlled environment (e.g. anechoic chamber). The second objective of this thesis was to examine whether VC is possible on ''wild'' data – VoxCeleb[2] dataset.
To evaluate effectiveness of the conversion, spoofing tests on speaker verification were conducted. Miss rate of verification system increased from 3.91 % up to 9.13 % which is relative change of **133.5 %**. These results also show importance of spoofing countermeasure development.

## Usage

Voice Conversion techniques can be used in various tasks. It is popular in funny internet videos but has also series of serious use cases, such as dubbing of audiovisual material and anonymization of voice (for example for witness protection). As it can serve for spoofing of voice identification systems, it is also an important tool for development spoofing detectors and counter-measures.

## References

[1] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, ''Autovc: Zero-shot voice style transfer with only autoencoder loss,'' *arXiv preprint arXiv:1905.05879*, 2019.

[2] A. Nagrani, J. S. Chung, and A. Zisserman, ''Voxceleb: a large-scale speaker identification dataset,'' *arXiv preprint arXiv:1706.08612*, 2017.