# 16 Pay attention to the speaker you want to listen to

## - Computational selective hearing based on deep learning -

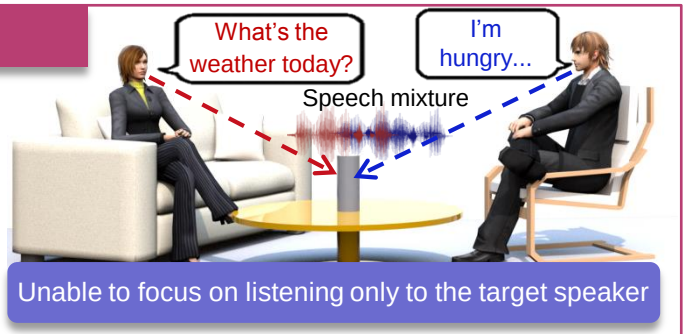
## Abstract

In conversations, when several people speak at the same time, people have the ability to focus on listening to a desired speaker (Selective hearing). However, current computers and voice assistant devices are not necessarily good at such hearing. We are pursuing research aimed at realizing computational selective hearing that will enable a computer to focus on listening to a target speaker and ignore the other speakers.

We use our recently developed context adaptive neural network and propose informing the neural network about the target speaker's voice characteristics such that the network can extract only that target speaker's voice. This technology will lead the way to a more natural voice assistant that can focus on listening to a target speaker in the same way that people do.

## Problem

- In everyday life, our words are often masked by someone else's voice e.g. in meetings or when a television is on in the background.
- For conventional voice assistant devices, it is hard to focus on a target speaker's voice when it is masked by others in a speech mixture.

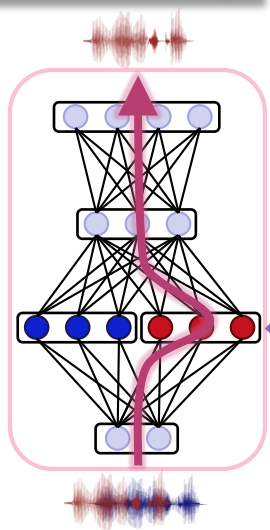



Picture designed with Sweet Home 3D. Includes 3D models created by Reallusion, Pencilart, Scopia and eTeks.

## Deep learning based selective hearing

We train a neural network to extract a target speaker's voice from a speech mixture, given the features of the target voice

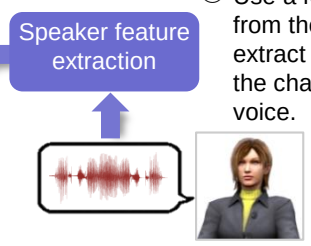
- ✓ By using a large amount of training data covering many mixtures of speakers, it is possible to extract target speakers unseen during training.
- ✓ Our proposed method does not require knowledge of the number of speakers in the mixture or the direction of the speakers, unlike conventional speech separation approaches.




③ Output speech of the target speaker only.

② Modify the behavior of the neural network according to the target speaker's characteristics (speaker features), to extract that target speaker.

① Use a few seconds of speech from the target speaker to extract features that represent the characteristics of his/her voice.




Picture designed with Sweet Home 3D. Includes 3D models created by Reallusion

## References

[1] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. Interspeech*, 2017.

[2] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, T. Nakatani, "Single channel speaker extraction and recognition with Speaker Beam," in *Proc. of 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'18)*, 2018.

## Contact

**Marc Delcroix** Signal Processing Research Group, Media Information Laboratory