A VISION-ASSISTED HEARING AID SYSTEM BASED ON DEEP LEARNING

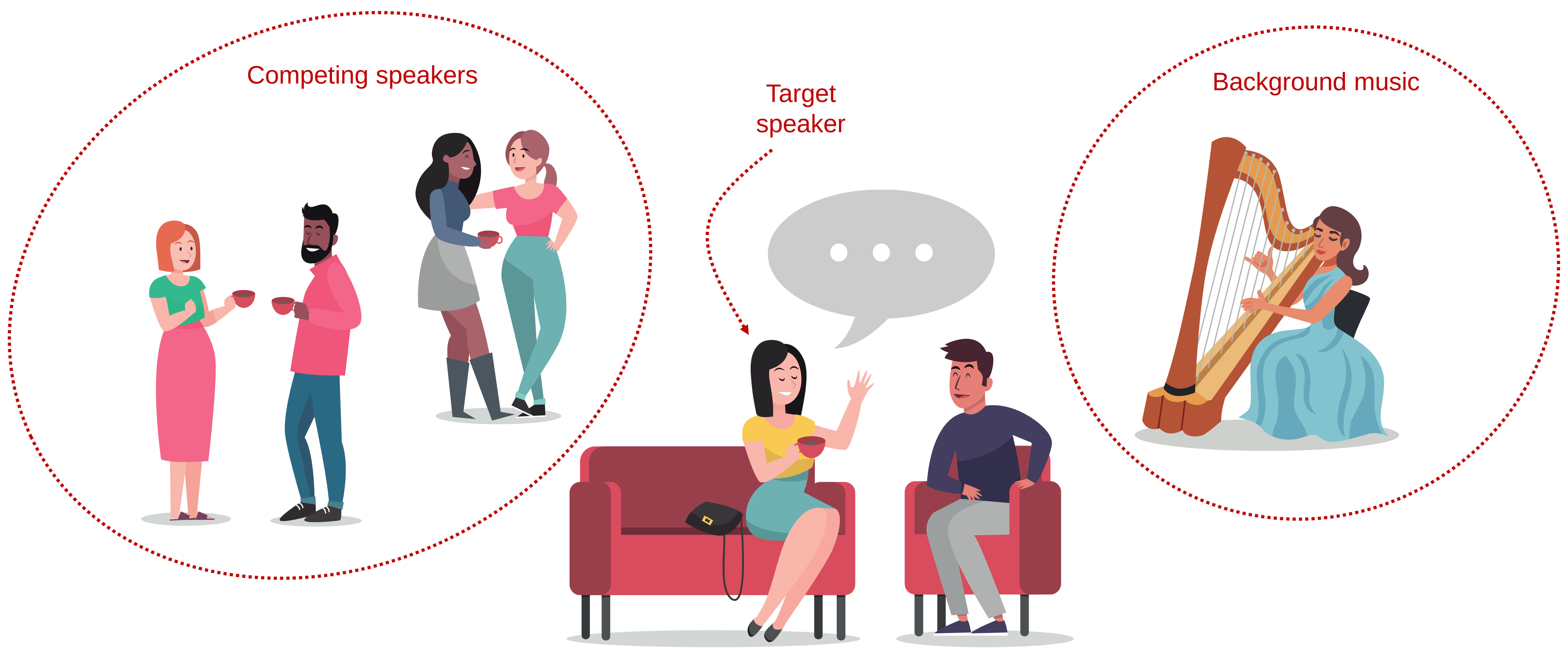
Daniel Michelsanti[1,2], Zheng-Hua Tan[2], Sergi Rotger Griful[3] and Jesper Jensen[1,2]

[1] Oticon A/S
[2] Aalborg University
[3] Eriksholm Research Centre

PROBLEM

Speech enhancement is the task of estimating the speech of a target speaker immersed in an acoustically noisy environment, where different sources of disturbance are present, e.g. competing speakers and background music.

Hearing aids can filter a noisy signal and provide an enhanced speech signal to the user. However, they perform poorly in particularly challenging noisy environments.



GOAL

Inspired by the human behavior in noisy environments, where visual cues are usually exploited in the form of **lip reading**, we want to improve the hearing aid performance using visual information.

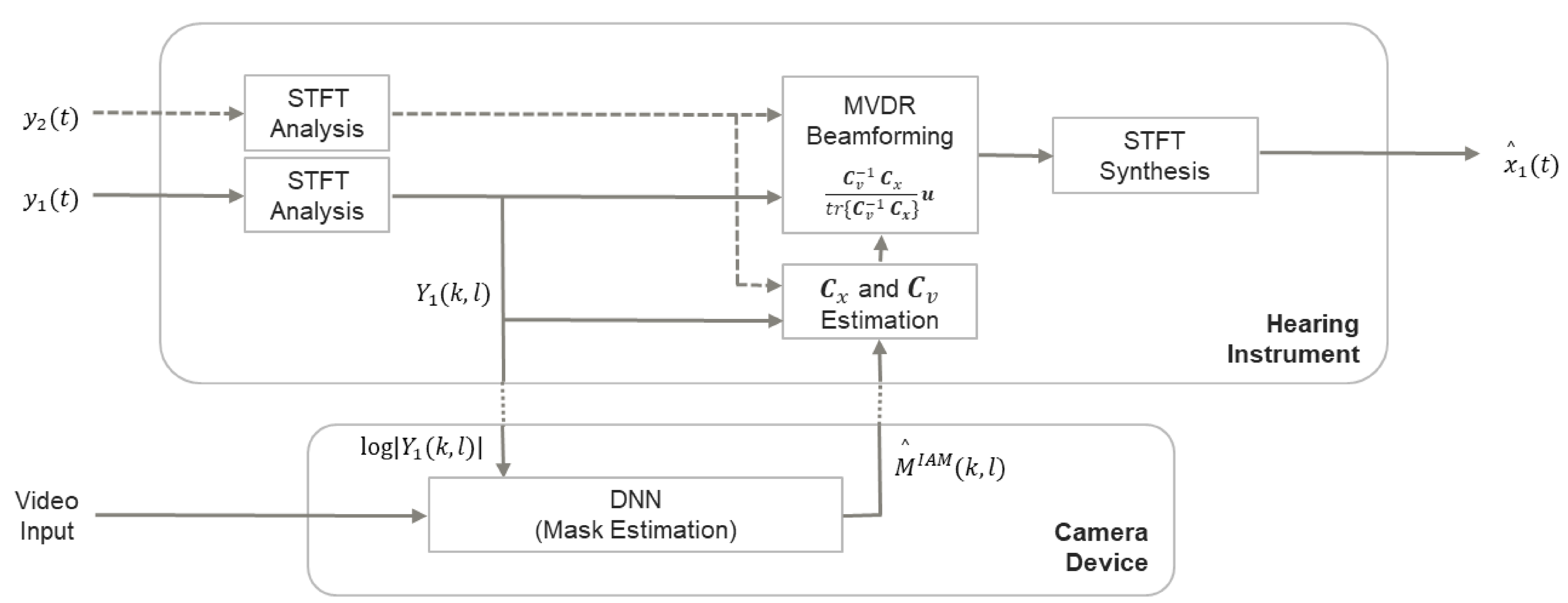
The idea is to use a device with a camera to enhance the performance of hearing aids in particularly challenging noisy environments.

METHODOLOGY

A **deep learning model** is trained to estimate a time-frequency mask from audio-visual data.

´The mask is used to estimate the inter-microphone power spectral densities (PSDs) of the clean and the noise signals.

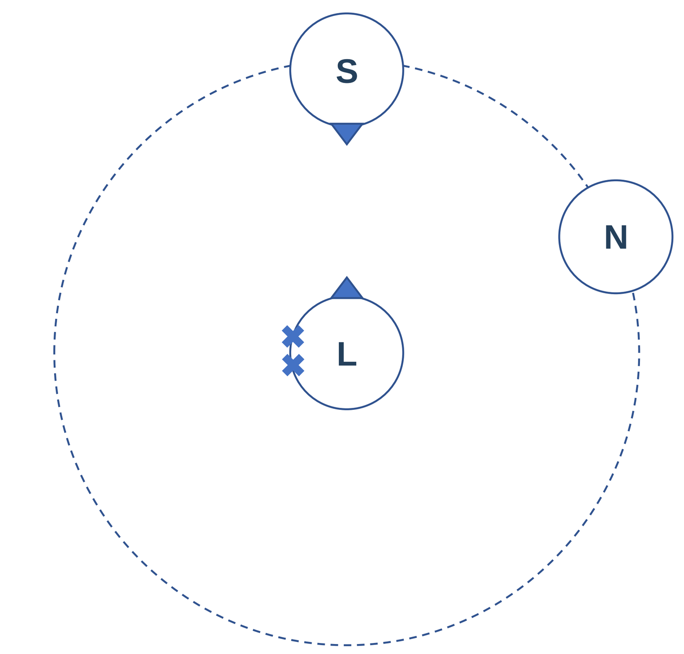
The PSDs are then used to build acoustic **MVDR beamformers**.



EXPERIMENTAL SETUP

Experiments are conducted on the **GRID dataset** in a speaker independent setting. A **2-channel hearing aid setup** is simulated with head-related impulse responses, in an anechoic setting.

We assume that the target speaker (S) is located in front of the hearing-impaired listener (L), while a point noise source (N) that generates white Gaussian noise at an SNR between -15 and 0 dB is located at 60 degrees.



RESULTS AND DISCUSSION

Results indicate that the multimicrophone audio-visual approach (AV MVDR) outperforms its audio-only multi-channel counterpart (AO MVDR) and single-microphone approaches (AO single-channel, AV single-channel) in terms of ESTOI and Segmental SNR.

As expected, the **biggest benefit** of our approach is **at low SNR** (-15 dB).