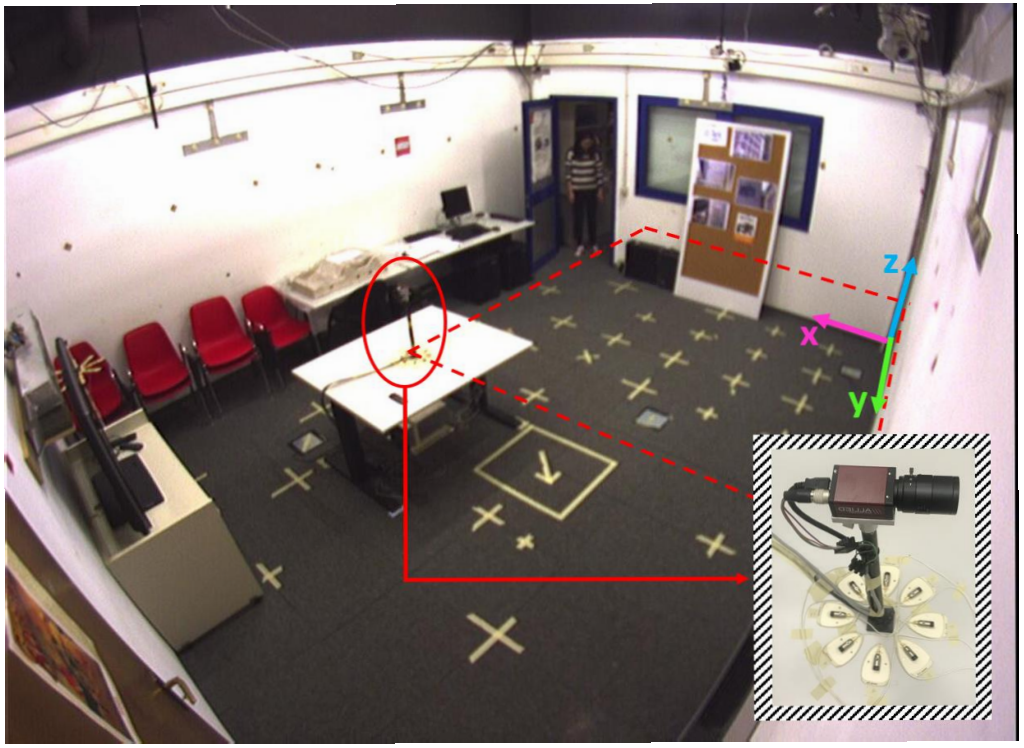
3D mouth tracking from a compact microphone array co-located with a camera

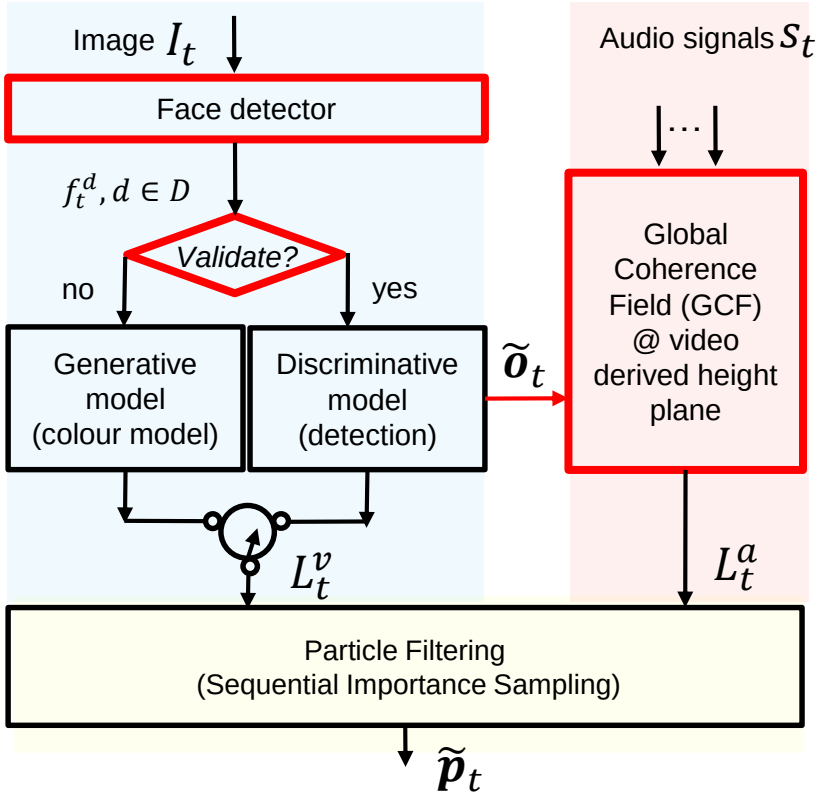
Xinyuan Qian

- Speaker tracking in 3D
- Audio-visual fusion
- Co-located sesing platform



Audio-Visual 3D tracker (AV3T)

- Face detection driven Particle Filter framework
- Visual likelihood: discriminative + generative models
- Audio likelihood: video-driven acoustic map



I_t Image at time index

f_t^d Face detection

D Set of face detections at time t

L_t^v Visual likelihood

$\tilde{\boldsymbol{o}}_t$ 3D mouth estimate

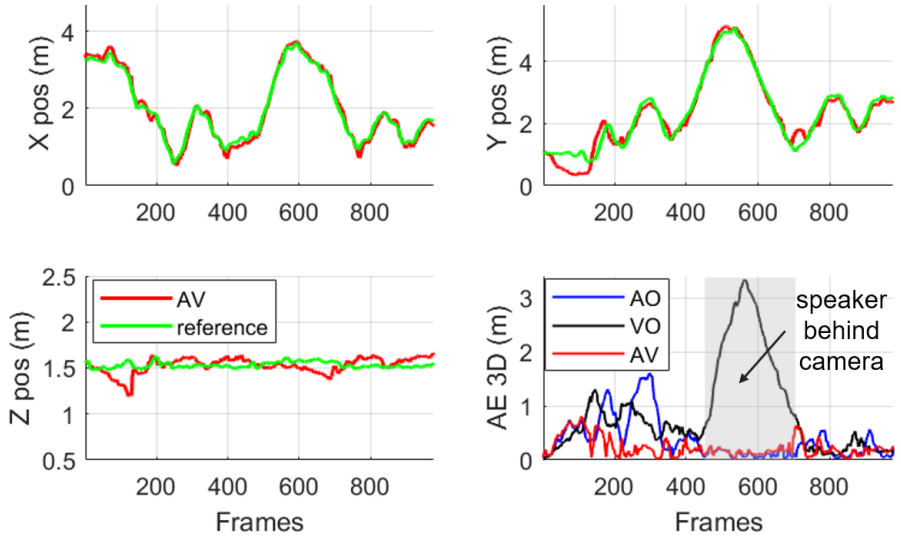
s_t Audio signals

L_t^a Audio likelihood

$\hat{\boldsymbol{p}}_t$ Target position estimate

Results and conclusion

- Superiority over uni-modal tracking results: e.g. audio-only (AO) and video-only (VO)
- An average 3D tracking accuracy of .25 m
- Capability of accurate 3D speaker tracking with co-located mutli-modal sensing platform



3D mouth tracking from a compact microphone array co-located with a camera
X. Qian, A. Xompero, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro
Proc. of IEEE Int. Conf. On Audio, Speech and Signal Processing (ICASSP), Calgary, Canada, 15-20 Apr 2018