Spatial Bias in Vision-Based Voice Activity Detection

K. STEFANOV, M. ADIBAN and G. SALVI

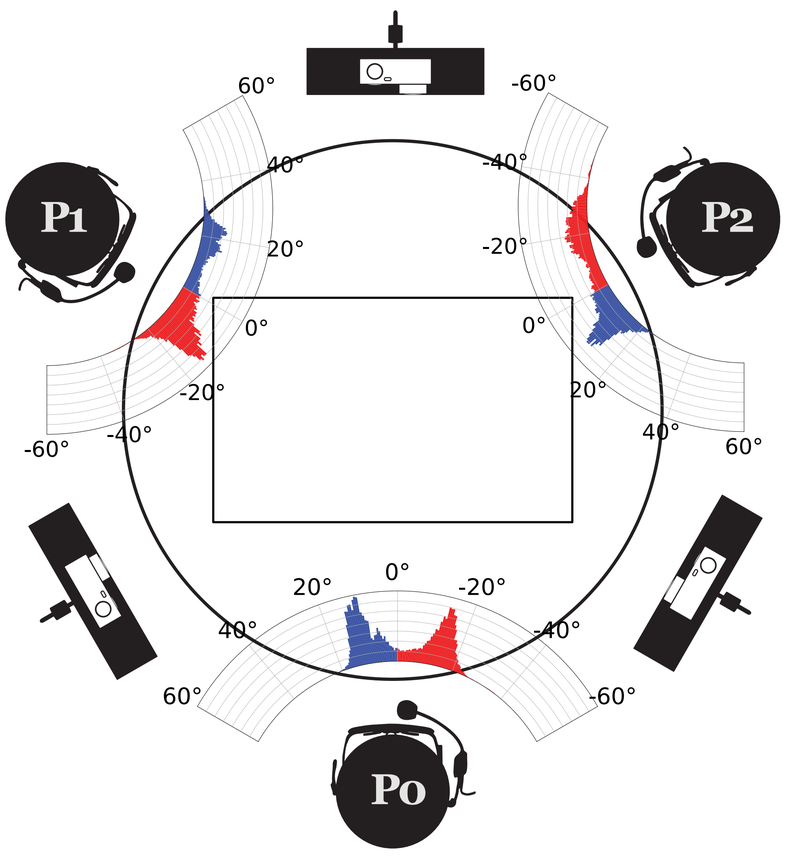
University of Southern California, Los Angeles, USA
Norwegian University of Science and Technology, Trondheim, Norway

kstefanov@ict.usc.edu, *mohammad.adiban@ntnu.no* and *giampiero.salvi@ntnu.no*

Contributions

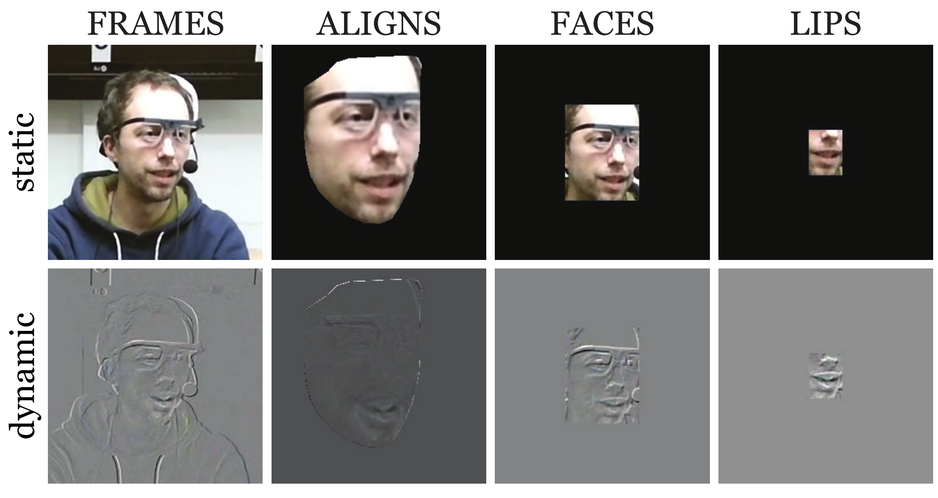
- We demonstrate that spatial bias (head pose information) related to the physical settings of the interaction is encoded in the learned representations of certain types of vision-based voice activity detectors
- We analyse the effect of data augmentation, input masking and dynamical inputs on the generalization capabilities of the models with mismatched train and test data
- We perform in-depth analysis of the features extracted by the models in order to explain the experimental results

Dataset

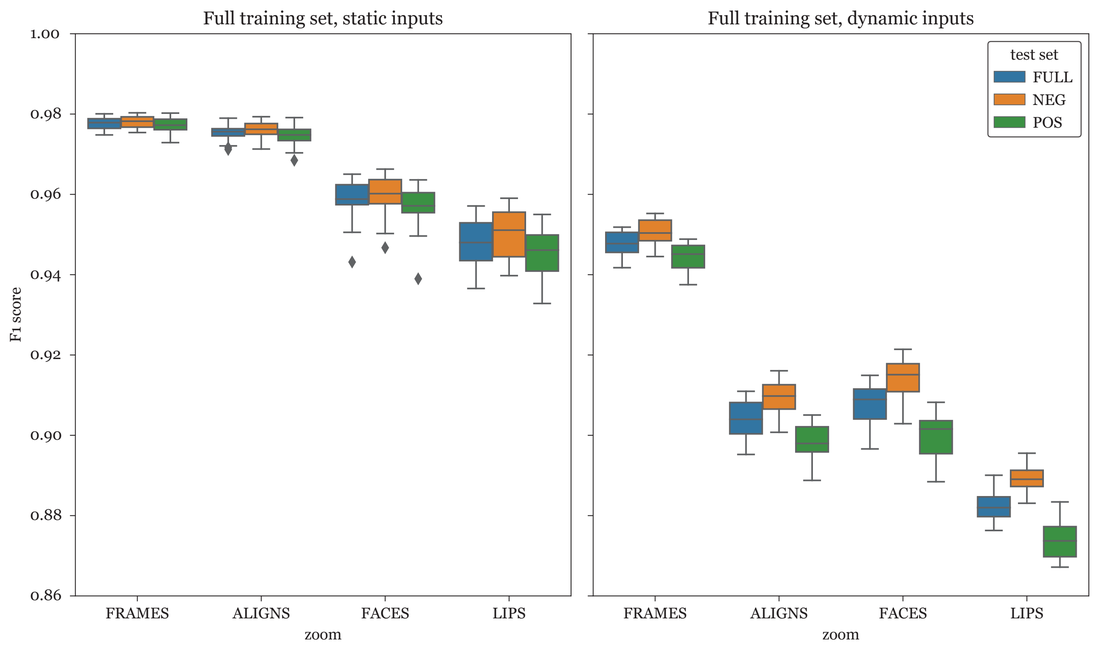


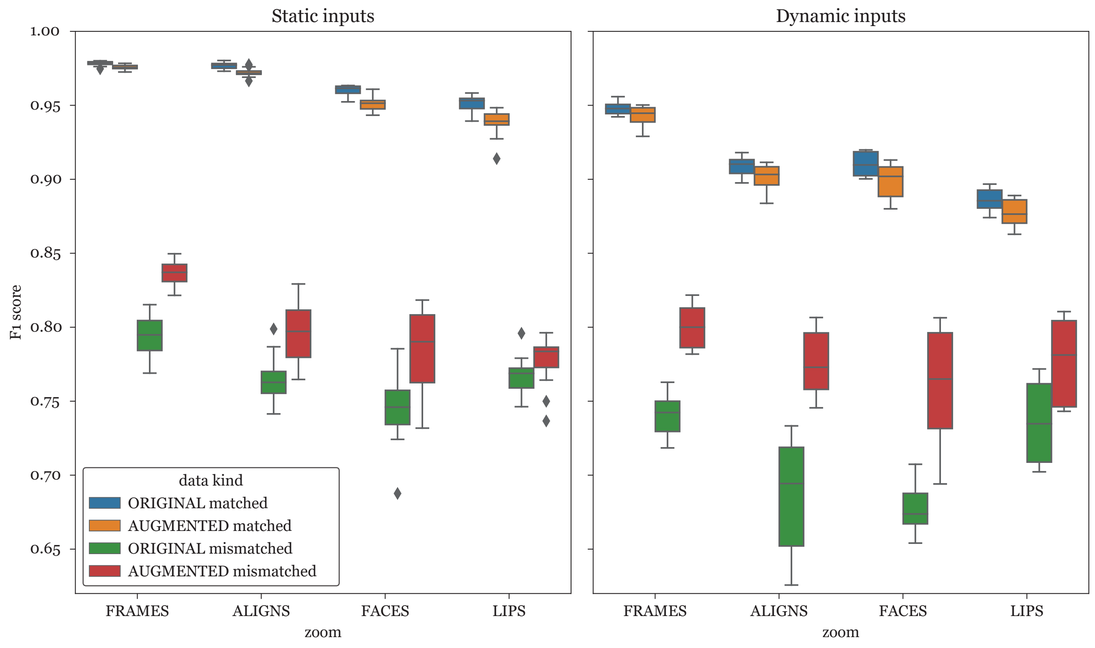
Models and Experiments

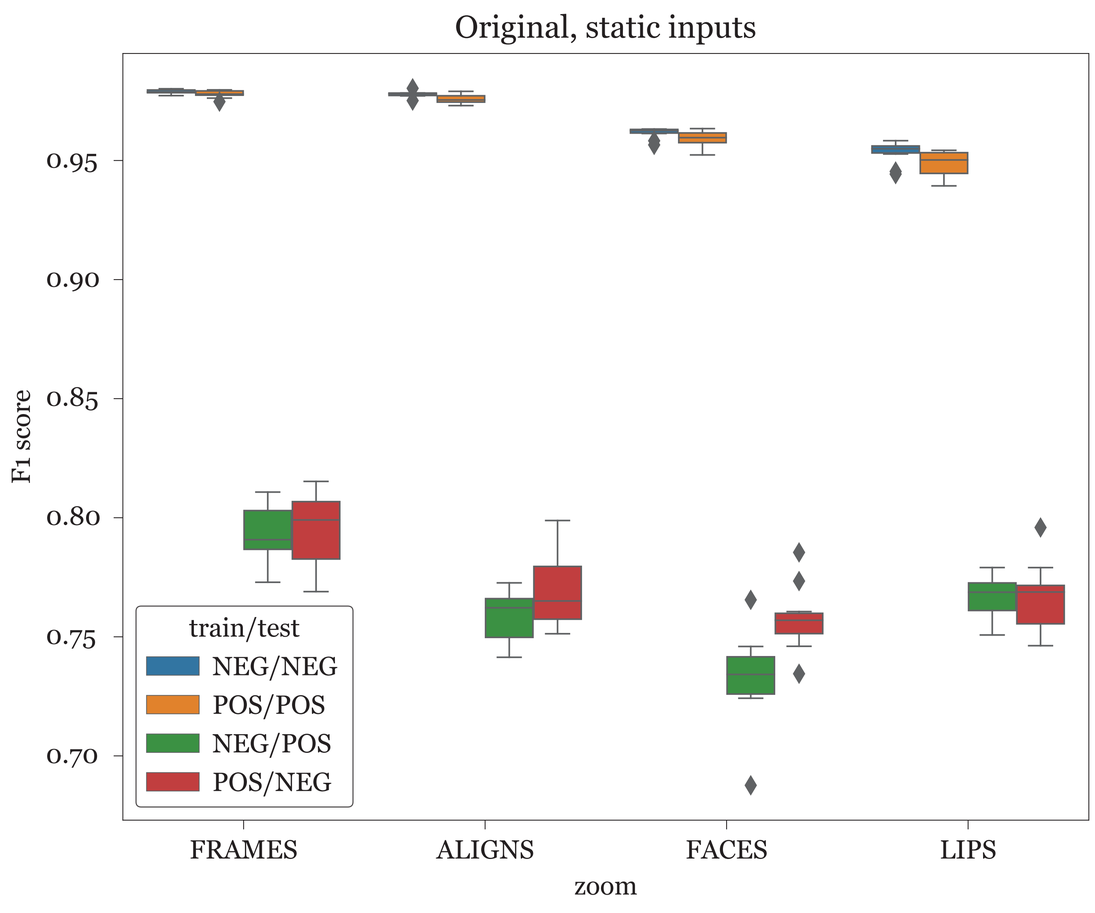
- ResNet-18 + classifier (simultaneously fine-tune and train)
- Masking (zoom) level (figure columns)
- Input type (figure rows)
- Data partitioning (positive and negative angles)
- Data augmentation (horizontal flip)



Results







This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.