Privacy-Preserving Representation for Audio-Visual Speech Understanding

Team Members: Raj Talan, Rajat Shrivastava, Shrey Agarwal

Problem Definition:
Multimodal datasets can contain personally identifiable information. We propose a general framework for privacy-aware representation of audio-visual (AV) data.

Relevance:
AV data is a core component of human communication and is the primary driver in the development of social intelligence. A general framework can ensure privacy across AV applications.

Datasets:
MSP-Improv: AV data of dyadic improvisation.
VidTIMIT: AV recordings of 43 people reciting short sentences.
VoxCeleb: Short clips of human speech extracted from interview videos

Planned Analysis:
We plan to adapt the architecture in **[2]** to incorporate AV data by:

1. Doing a literature review on privacy transformers
2. Implementing AV-HuBERT **[3]** as an encoder
3. Developing a Privacy Filter for a multimodal input stream
4. Comparing with existing baselines

Expected Outcome:
A framework to create privacy-aware representations of AV data that can be used across various applications, with results that are on par with current baselines.

References:

1. H. Xu, Z. Cai, D. Takabi and W. Li, "Audio-Visual Autoencoding for Privacy-Preserving Video Streaming," in IEEE Internet of Things Journal, vol. 9, no. 3, pp. 1749-1761, 1 Feb.1, 2022, doi: 10.1109/JIOT.2021.3089080.
2. Tran, Minh, and Mohammad Soleymani. "Privacy-preserving Representation Learning for Speech Understanding." INTERSPEECH 2023.
3. Shi, B., Hsu, W.N., Lakhotia, K. and Mohamed, A., 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184.