German Lip-Reading System

Zhengyang Li, Thomas Graave, Matthias Dunkelberg
Mentor: Jing Liu

German Lip-Reading System
Motivation

- **Lip-Reading** utilize lip move rances
- **Application**: transcribing archival silent films [1], helping people that suffer from aphonia, or supporting in crime investigations.
- **Challenge:** The **lack of labeled German audiovisual data** hinders the development of German audiovisual speech recognition systems
- **Method: Transfer learning** from high-resource languages is an efficient way to improve the audiovisual ASR for low-resource languages
- **Task: German lip-reading system** and the implementation of an **interactive demonstrator**

German Lip-Reading System
Method - Transfer Leaning

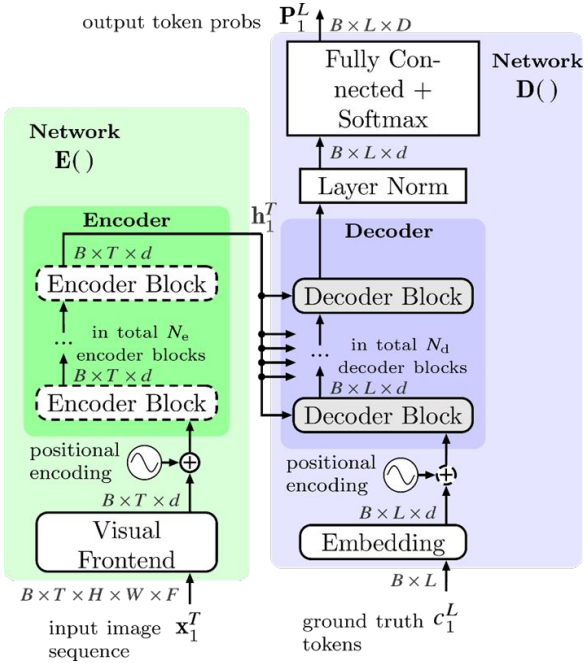


Figure 1. transformer encoder-decoder neural network for audiovisual speech recognition

- **Source Model and Data**: an **AV-HuBERT** Model [Shi,ICLR,2022] trained on **English** audiovisual data.
  - The training of an AV-HuBERT model consists of two stages:
    - Self-supervised pre-training of the **encoder network** on **unlabeled** data: $\mathcal{D}_{S=VoxCeleb2}$ $\mathcal{D}_{S=LRS3}$
    - Supervised fine-tuning of the **encoder-decoder** network on **labeled** data: $\mathcal{D}_{S=LRS3}^{\star}$
- **Target Model and Data**: A **German** audiovisual ASR trained on labeled German audiovisual data: $\mathcal{D}_{T=De}^{\star}$

$\star$: Labels are available (supervised learning)

[Shi,ICLR,2022] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction,” in Proc. of ICLR, virtual, Apr. 2022, pp. 1–24.

German Lip-Reading System
Datasets

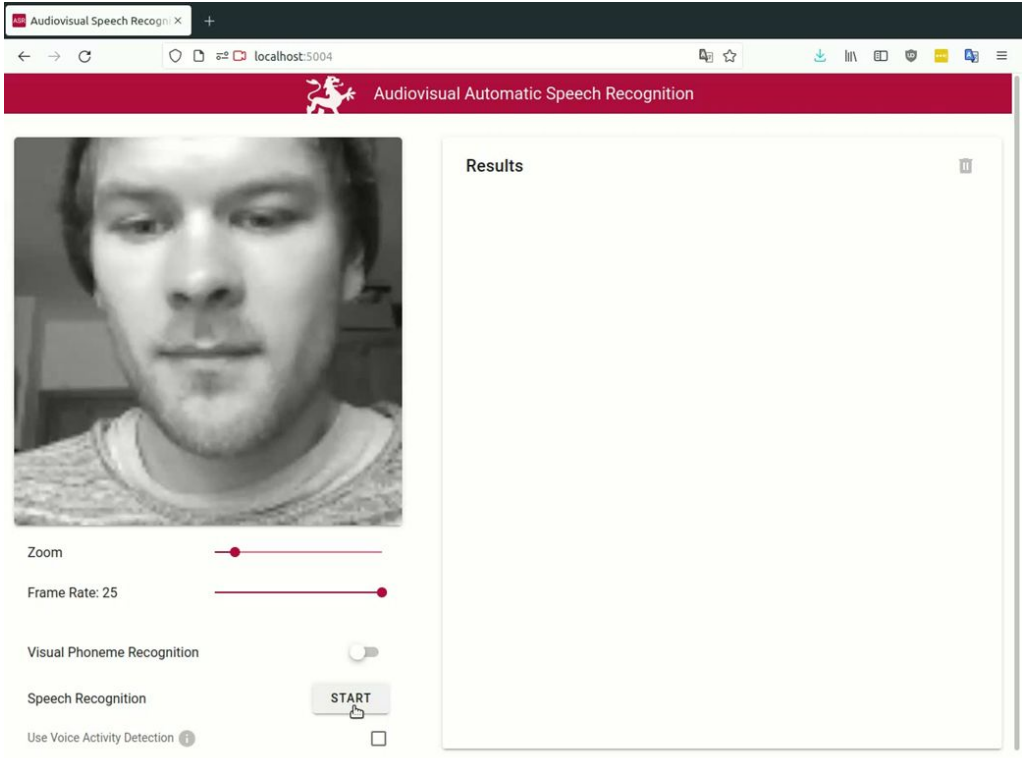
Table 1: Audiovisual Datasets in English and in German

	Datasets	Language	Size	Type	Description
Source data	LRS3	English	433h	Sentence-level audiovisual datasets	• Audio:16kHz, Video: 25fps
Source data	VoxCeleb2	Multilingual	2,442h (1,326h in English)	Sentence-level **Multilingual** audiovisual datasets **without** text annotations	• Audio:16kHz, Video: 25fps • Only the English data is used to pre-train AV-HuBERT • No language label for utterances
Target data	SmartKOM	German	30h	Sentence-level	• Audio:16kHz, Video: 25fps

- **Training:** The lip-reading model is **pre-trained** on LRS3 and VoxCeleb2 unlabeled English audiovisual data and **fine-tuned** on SmartKOM German video data.

https://www.dfki.de/web/forschung/projekte-publikationen/projekt/smartkom/

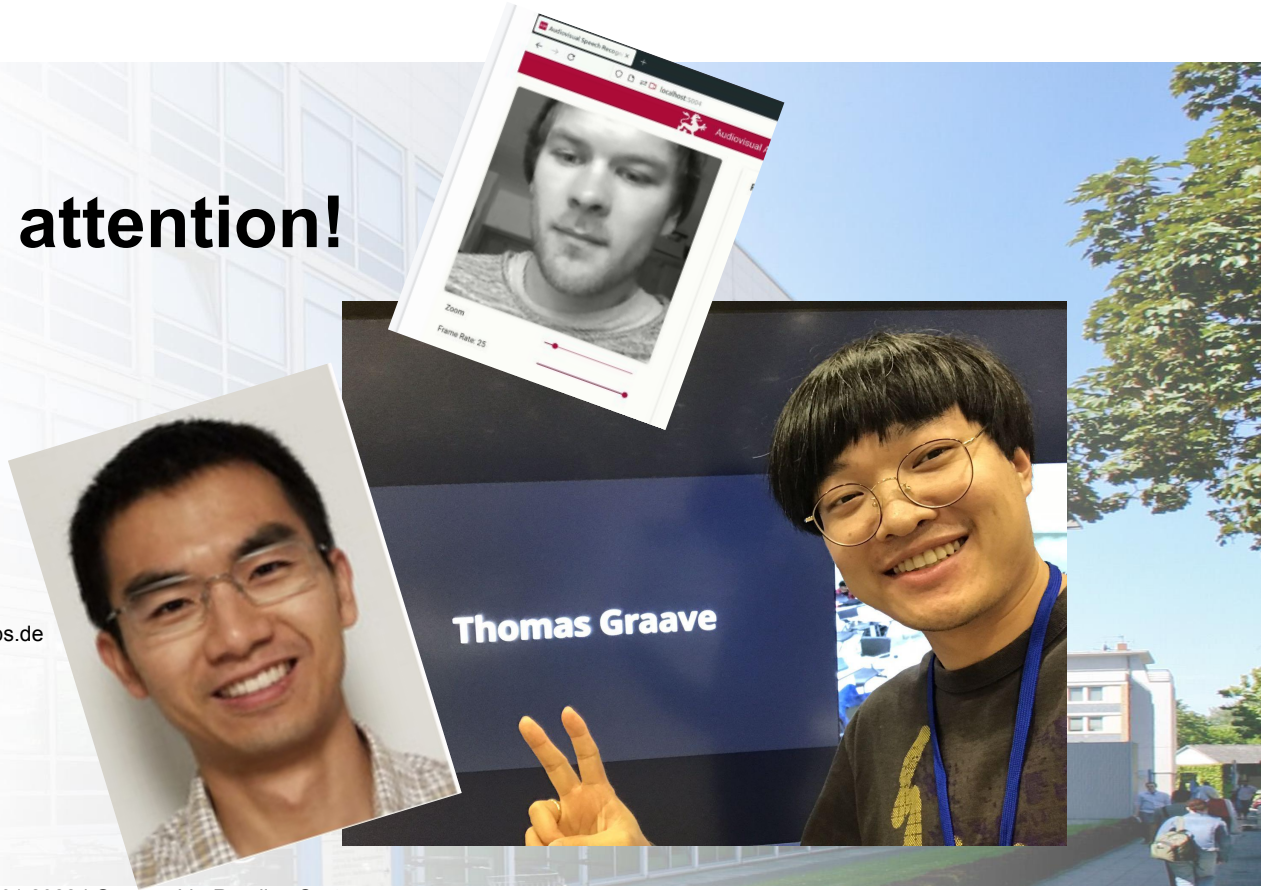
German Lip-Reading System
Results



- **Demonstrator:**
  - **Face Tracking**
  - **Extraction of ROI**
  - **Lip-reading and audiovisual ASR**

Methods	WER (%)
	valid
Lip-reading	68% (3000/10000updates)

Thanks for your attention!



Zhengyang Li
Thomas Graave
Matthias Dunkelberg
{zhengyang.li, thomas.graave, matthias.dunkelberg}@tu-bs.de

Mentor: Jing Liu
jlmk@amazon.com