Week 3

Sunday, September 28, 2025 4:03 PM

Overview of the week

AV-Hubert set up in colab

- Models downloaded and uploaded
- All issues with libraries and version compatibility resolved
- Layout of directory setup
  - Keep tidy and don't have tests and results get mixed up
- Logic to freeze environment so it can be replicated at a later date

Papers:

- Look into models and techniques for sign language to help with ideas for visemes
- Techniques that keep audio and visual sides same length
- Read AV-Hubert papers
- How much information can be got from audio and visual (certain phonemes and visemes)
- What frame rate is needed to capture mouth movements

Augmentation Ideas

- Upsample/Downsample
  - Interpolation (inbetweening) using open source models to boost from 30fps to 60
  - Shown that lip reading models struggle when validation data at a different fps than its training data
    - Ex. Trained on 30fps and validated with some 60fps av data
    - Could be an opportunity to make a model more robust
  - Data could be upsampled and then downsampled back to 30fps with different sampling phases
    - Ex. Upsample with interpolation and then downsample back to 30fps by taking every first frame or every second frame
  - Things to watch out for:
    - Keep audio and visual in sync and same length
    - Test interpolation models because artefacts could cause data to do more bad than good

- **ROI constrained Erasing**
  - Intentionally mask/blur specific mouth regions during training to prevent over reliance on a single visual cue. Lots of phonemes can map to the same viseme but different phonemes carry different amounts of information
    - Let's say there are three key parts of the mouth that move to create the /g/ sound, if you erase 1 out of 3 or 2 out of 3 does this improve robustness
    - Maybe it's important to become accustom with the key mouth and jaw areas and how they map to certain visemes depending on their action, if they are blocked this could make it more difficult for the model to purely rely on that mouth region and its given action

- **Using DeepFakes**
  - FaceSwap,Fsgan, wav2lip
  - Datasets exist –
    - FakeAVCeleb
      - https://sites.google.com/view/fakeavcelebdash-lab/download?authuser=0
    - DeepfakeTIMIT
    - UADVF
    - FaceForensics++
    - (FF++)
    - Celeb-DF
    - Google DFD
  - This paper outlines harnessing av-hubert to extract visual and acoustic features and then designed their own model to detect deepfakes – what if deepfakes used to train av-hubert
    - https://arxiv.org/pdf/2311.02733

Best way to access dataset

- To prepare my first baseline test in colab what is the best way to access the TCD-TIMIT dataset
  - I requested access through the sigmedia website
  - Is this the best way to go about this?