**Speaker identification.** involves determining which speaker has produced a given utterance, if this is performed for a closed set of speakers then the task is similar to that of multiclass classification.

**Speaker verification.** on the other hand involves determining whether there is a match between a given utterance and a target model.

**Cross-modal verification.** To determine whether two inputs from different modalities are semantically aligned. More specifically, given a face input and a speech segment, the goal is to determine if they belong to the same identity.

Defining benchmarks for task: Two evaluation protocols

Seen-heard

Unseen-unheard

Computer Vision Task

1. Image Obfuscation
2. Weaker authenticators
3. Joint Embedding Tasks

Obfuscated data creation task

Little pre-processing

VGG-Vox architecture

Speaker Verification

Input: Short-term amplitude spectrogram, extracted from 3 seconds of raw audio using a 512-point FFT

At training time, 3s segment of audio is chosen randomly from entire audio segment.

Dimensionality of final FC layer is reduced to 256, 256-D embedding are L2-normalised.

# epocs 50

Batch size = 256

SGD momentum 0.5

Weight decay 5e-4

Log decaying learning rate (10-2 to 10-8)

Contrastive loss (0.2, 0.4, 0.6, 0.8) - not done yet

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Verification split**   |  |  |  | | --- | --- | --- | |  | **dev** | **test** | | **# of speakers** | 1,211 | 40 | | **# of videos** | 21,819 | 677 | | **# of utterances** | 148,642 | 4,874 | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train** | **Test (S-H)** | **Val(US-UH)** | **Test(US-UH)** |
| **# of identities** | 901 | 901 | 100 | 250 |
| **# of videos** |  |  |  |  |
| **# of utterances** |  |  |  |  |

Our goal is to determine whether there is a match between a given utterance and a target model.