**AUDIO VISUAL SPEECH RECOGNITION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Dataset name | Characteristics | Techniques | Publisher name and year of publishing |
| 1. | **VoxCeleb2**  Deep Speaker Recognition | * Large-scale speaker recognition dataset obtained automatically from open- source media. * Contains over 6,000 speakers * 1 million utterances * Fairly gender balanced. 61% male speakers. * Collected “in the wild”. * Multilingual, with speech from 145 different nationalities. * Surpasses Voxceleb1 | * Convolutional Neural Networks. * ResNet- based architectures. * Automatic computer vision pipeline | Joon Son Chung, Arsha Nagarani, Andrew Zisserman.  (2018) |
| 2. | **LRS3 – TED**  LRS3-TED : a large scale dataset for visual speech recognition  Audio-Visual speech recognition is worth 32x32x8  voxels | * Multi – modal dataset for visual and audio-visual speech recognition. * Includes face tracks over 400 hrs , extracted from 5594 TED and TEDX, with subtitles and word alignment boundaries. * Cropped face tracks are provided as .mp4 files * Resolution of cropped face tracks : 224 X 224.   Frame rate : 25 fps   * Single channel 16- bit 16kHz format audio tracks are provided. * Corresponding text transcripts , alignment boundaries of every word included * Dataset statistics: * Consists of recordings of TED talks. * LRS3-TED eval set is smaller and simpler than YTDEV18 * Audio quality is high and the video is clean, high definition, and almost always centred at the speaker. * Audio-visual automatic speech recognition (AV-ASR) uses LRS3 dataset for training its system. | * Convolutional neural networks * Multi-stage pipeline * Convolutional neural networks. | Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman.  (2018)  Dimitriy Serdyuk, Otavio Braga, Olivier Siohan.  (2021) |
| 3. | **LRS2 - BBC**  Deep audio-visual speech recognition | * Formed by collecting and pr-processing thousands of videos from the British television * Consists of talking faces together with subtitles. * The videos contain faces “in the wild”. * Two models are considered that can transcribe audio and video sequences of speech into characters. * Same architectures can be used when one of the modalities is present. * In presence of noise in the audio, combining two modalities lead to improvement. * Dataset statistics: | * Convolutional neural networks. * Connectionist Temporal Classification (CTC) loss * Sequence -to- sequence (seq2seq) loss * Multi-stage pipeline | Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman.  (2018) |