**Why a new audio-visual corpus?**

The broadness and wide coverage of existing multi-speaker audio-visual corpora like the LRS-2 and LRS-3 comes with some drawbacks—

1. Very little material is available per speaker and snippets tend to be short and typically contain only single phrases.  
   > This prohibits modelling long-range phenomena like conversational prosody.
2. They lack multilingual data from the same speaker.  
   > These are required to evaluate cross-lingual machine learning tasks such as lip-synchronous audio-video translation.

**How does our corpus resolve these?**

The **Merkel Podcast Corpus** is an audio-visual-text corpus in German collected from 16 years of (almost) weekly internet podcasts of former German chancellor Angela Merkel. This corpus—

1. contains large amounts of speech from a public figure, to the extent that it can meaningfully be used to train modern single-speaker deep learning models;  
2. also contains further material from many other speakers (interviewers) which can help generalize to multi-speaker models;  
3. contains material that is both closely aligned (short snippets of video with the corresponding texts) and consecutive in nature, i.e., snippets can be contextualized into the overall situation;  
4. comes with large amounts of meta-data available, such as date of recording, speaker names and further written documentation that the speech material can be related to;  
5. is amended with the few publicly available video recordings of Angela Merkel speaking English, which is particularly useful for cross-lingual processing tasks such as evaluating lip-synchronous dubbing.

**Dataset Statistics**

- The overall corpus consists of 630 videos totaling 48 hrs. of material in German:  
  - 2.7 hrs. of leading and trailing jingles and 2.8 hrs. of pauses  
  - 42.5 hrs. of speech  
- More than 250 interviews with different interviewers, providing rich multi-speaker background.  
- Audio is encoded as high quality AC3 and high quality video (HD since 2018)  
- Angela Merkel appears on-screen 66% of the time and is the on-screen active speaker for 58% of the time.  
- We curate a single speaker corpus whose snippets have mean duration of 7.2 s (stdev: 6.2 s) and mean text length of 124 characters (stdev: 104 characters).  
- We manually align English utterances spoken by Merkel which amount to 51 snippets and 256 seconds of speech.

**Machine-Learning Applications**

- **Age Estimation**  
  - Using softmax logistic regression using speaker embeddings, to estimate the year of recording yields macro f1 : 0.63 (regression coefficient : 0.77).  
  - This indicates speaker voice change with age of speaker and/or recording.
- **Lip Generation**  
  - We train Wav2Lip (speech-to-lip generation) using single-speaker corpus.  
  - Lip renderings in videos look more natural and appear to better match Merkel’s behavior (She tends to not open her mouth as far as off-the-shelf Wav2Lip model).  
- **Visually Grounded Speech Synthesis**  
  - We train a visually grounded TTS system on the single-speaker corpus.  
  - Often the synchrony improves drastically. But long snippets, prosodic structures and pauses sometimes lead to erratic results, requiring further investigation.

**Conclusions**

We have presented a multi-modal corpus with one primary speaker and more than 250 secondary speakers spanning over a period of 16 years. We supplement this corpus with English speech of the primary speaker with hope that this will aid research in lip-synchronous audio-video translation.

https://github.com/deepslsd/Merkel-Podcast-Corpus

**Contact**

Timo Baumann  
OTH Regensburg  
Email: mail@timobaumann.de  
Website: https://timobaumann.de/