Merkel Podcast Corpus

A Multimodal Dataset

Compiled from 16 Years of Angela Merkel's Weekly Video Podcasts

Debjoy Saha[1,2], Shravan Nayak[1,3], Timo Baumann[1,4]
[1]Universität Hamburg, [2]IIT Kharagpur, [3]IIT (BHU) Varanasi, [4]OTH Regensburg

LREC 2022 Marseille

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.

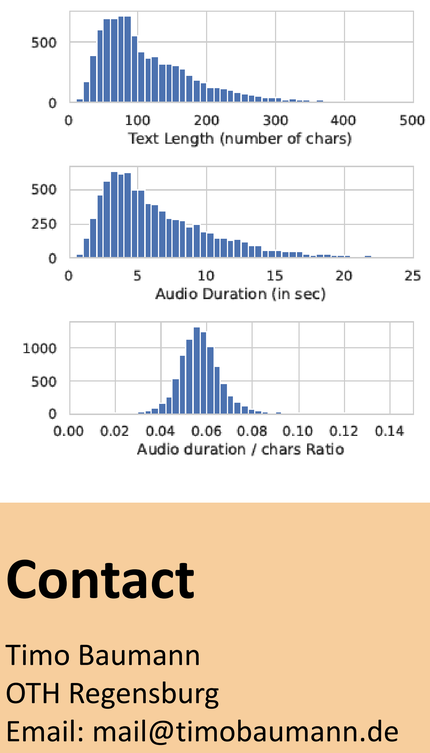
How does our corpus resolve these?

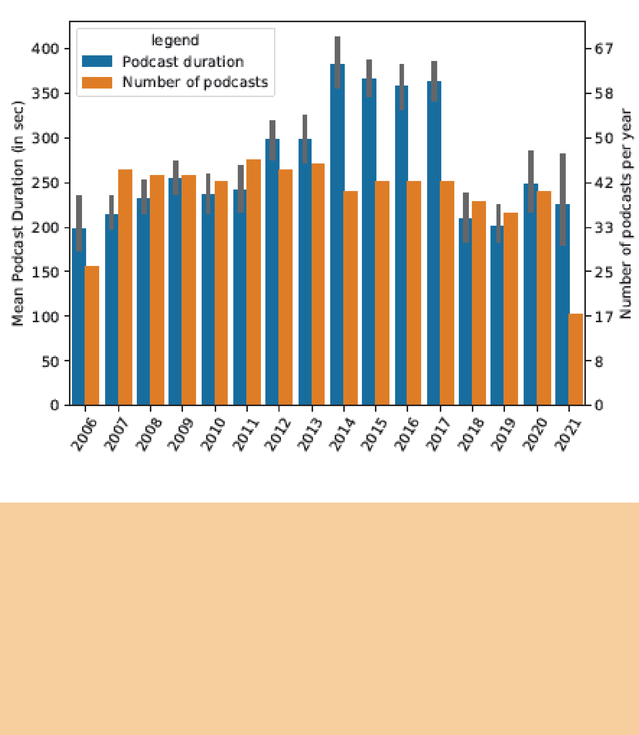
The **Merkel Podcast Corpus** is an audio-visual-text corpus in German collected from 16 years of (almost) weekly internet podcasts of former German chancellor Angela Merkel. This corpus -

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.





Dataset Source and Preparation

Scraping: More than 600 such podcasts (videos, subtitles and metadata), published between 2006 and 2021, are scraped from publicly available archives[1]. Since 2013, podcasts contain lightly edited subtitles; older videos are amended with a written transcript. While originally recorded in the form of semi-spontaneous speeches, the format changes to interviews post-2011.

Alignment and Snippeting: Due to the absence of subtitles pre-2013, we use robust forced alignment[2] to align text and speech. Subsequent cropping to smaller snippets is done on the basis of speech pauses.

Speaker Diarization: Single-speaker audio-visual corpora should contain video where the face of the target person is visible and she is actually speaking.

- We temporally crop snippets into one or multiple scenes (cropping at discontinuities like camera changes) using a scene-detect tool[3].
- We then crop the faces of the people identified in each scene[5].
- Talknet tool[4] can be used to detect whether the speaker on-screen is talking.
- A Face recognition tool[5] is used to check similarity with a reference set of images of the target speaker.

> A positive result in both of these tests ensures that our target speaker is speaking on-screen in the face-cropped scene. If for some scene, none of the cropped faces passes the above two checks, the entire snippet is discarded.

Figure 1. Speaker Diarization Pipeline for semi-automatically differentiating speakers and tuning in on one target speaker in collections of multi-speaker videos based on a few examples of the target speaker.

Machine-Learning Applications

- **Age Estimation**
  - Using softmax logistic regression using speaker embeddings[6], 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/deeplsd/Merkel-Podcast-Corpus

Contact

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

References

1. https://www.bundesregierung.de/breg-de/service/archiv/archiv-podacasts
2. https://link.springer.com/article/10.1007/s10579-017-9410-y
3. https://github.com/Breakthrough/PySceneDetect
4. https://github.com/TaoRuijie/TalkNet_ASD
5. https://github.com/ageitgey/face_recognition
6. https://github.com/RF5/simple-speaker-embedding