Short description of the

Vera am Mittag German Audio-Visual Spontaneous Speech Database

"VAM Corpus"

The audio-visual speech corpus proposed for distribution by LDC was collected by the Institut für Nachrichtentechnik of the Universität Karlsruhe (TH), Karlsruhe, Germany, for research purposes. This data collection was supported by grants of the Collaborative Research Project (SFB) 588 "Humanoid Robots – Learning and co-operating multimodal robots" by the Deutsche Forschungsgemeinschaft. The data is provided by the German producer *time 2 talk entertainment GmbH* , Potsdam , Germany.

The VAM corpus consists of 12 hours of recordings of the German TV talk-show "Vera am Mittag" (Vera at noon), which are segmented into broadcasts, dialogue acts and utterances, respectively. This audio -visual speech corpus contains spontaneous and very emotional speech recorded from unscripted, authentic discussions between the guests of the talk-show. Such data is of great interest to all research groups working on spontaneous speech analysis, emotion recognition in both, speech and facial expression, natural language understanding, and robust speech recognition. Further interests may arise from a linguist's viewpoint in the variety of German regional accents that are present in the data.

In addition to the audio-visual data and the segmented utterances we provide emotion labels for a great part of the data. This labeling follows state-of-the art insights from emotion psychology. Thus, the emotion labels are given on a continuous-valued scale for three emotion primitives: valence (positive vs. negative), activation (calm vs. excited) and dominance (weak vs. strong). , using a large number of human evaluators.

The data is structured as follows:

1. VAM-Video
   This part of the corpus contains the audio-visual signals structured as 12 broadcasts (shows). 10 of the broadcasts are sub-structured as 4-5 dialogs, the remaining two broadcasts are only added in total as mpg files. Each dialog is provided as one audio-visual mpg file. In addition, each dialog is also segmented into utterances, and for each utterance, three data sources are provided: the audio-visual speech signal as a mpg file, the audio-only signal as a wav file, and the visual signal only as a sequence of png files. In total, the VAM-Video part contains 1421 segmented utterances of 104 different speakers. Due to the video files on both sentence level and dialog level (which is necessary to allow for the use of timing information), the total disk size sums up to 19,0 GB.
   For this part, no emotion labels are available. The next two corpus parts are extracted from the VAM-Video corpus.

2. VAM-Audio
   This part of the corpus contains the audio signal only, and the emotion labels given by several independent human evaluators. The data is organized speaker-wise for 47 individual speakers. For each speaker, the data is sub-structured in sentences, and for each sentence, we provide the wav file, the emotion labels given by 17 (speakers 1-19) or 6 (speakers 20-47) human evaluators, and the fused emotion evaluation results.
   In total, 1018 utterances are contained in the VAM-Audio corpus. To distinguish between the first set of recordings which was emotion-labeled by a larger number of evaluators and the second set of recordings, we suggest to call the utterances by speakers 1-19 VAM-Audio I and the rest VAM-Audio II. The corpus size of VAM-Audio is 177 MB.

3. VAM-Faces
   This part of the corpus contains extracted facial images of the speakers in the VAM-Video corpus. The data is organized speaker-wise for a reasonable subset of 20 speakers. For each speaker, the data is sub-structured in sentences. For each sentence we provide the audio-visual signal as a reference, and in addition the facial image as a png file as well as a emotion category label and emotion primitive labels for many (not all) frames of the sentence.
   The corpus VAM-Faces therefore contains 1872 emotion-labeled facial images extracted from audio-visual speech recordings of the VAM-Video corpus. The corpus size on disk is 600 MB.

The wav files are provided at 16 kHz sampling rate and 16 bit resolution as stereo signals. The png images have a resolution of 352x288 pixels. The video files are MPEG-coded image sequences of the same resolution with a frame rate of 25 fps, and the audio stream sampled at 44.1 kHz, stereo.

For further information, please contact:

Dipl.-Ing. Michael Grimm
Universitaet Karlsruhe (TH)
Institut fuer Nachrichtentechnik (INT)
Kaiserstr. 12
76128 Karlsruhe, Germany
Tel: ++49+721-608-3790
Fax: ++49+721-608-3799
E-Mail: grimm@int.uni-karlsruhe.de