AI-SPEAK database

The dataset comprises recordings from 30 speakers, including 15 female and 15 male participants. Prior to preprocessing, each speaker contributed 160 recordings: 80 in Serbian and 80 in English. Out of the 80 sentences per language, 30 are shared across all speakers, while the remaining 50 are speaker-specific (personalized). Each speaker was recorded using a professional microphone (audio only) and three video cameras—one high-quality front-facing (frontal) camera and two auxiliary mobile phone cameras positioned approximately 30 degrees to the left and right.

The dataset underwent manual inspection and processing. The final version is publicly available and organized as follows:

Each ZIP archive contains five folders named *spkXX*, where *XX* refers to the speaker ID (ranging from 01 to 30). Each folder holds data corresponding to a single speaker and contains three subfolders: *alignment*, *ser*, and *eng*, along with an *Excel* file. The *Excel* file contains eight columns. The first column (*name*) specifies the filename. Columns two to five (*video_a, video_r, video_l, audio*) indicate the availability of recordings from the frontal, right, and left cameras, and the microphone, respectively. The sixth column (*transcript*) contains the spoken utterance corresponding to the recordings. The seventh column (*language*) indicates the language of the utterance, labeled as *ser* for Serbian or *eng* for English. The final column (*common*) specifies whether the sentence is part of the common subset shared across all speakers (*true*) or part of the speaker's personalized subset (*false*).

The alignment folder contains *.align* files, one per audio recording, with word-level time alignments in milliseconds specifying the start and end time of each word. The *ser* and *eng* folders contain language-specific materials and are structured as follows.

Each contains four subfolders: *video_a_anonymized, video_r_anonymized, video_l_anonymized*, and *audio*. Within each *video_x_anonymized* folder (where x denotes camera angle), there are video recordings. Video recordings contain lip-only video clips, where the speaker's lips are clearly visible while the area around it is pixelized (in order to be anonymized but still key points of the face are detectable), while the remainder of the frame is masked (black). The *audio* folder contains the corresponding audio recordings.

Notes:

- All recordings with the same filename (from the *audio, video_a, video_l,* and *video_r* folders) are time-synchronized, i.e., they share the same transcript and alignment information.

- Alignments were generated automatically and may contain occasional inaccuracies; this component was not manually verified.
- Recordings may contain brief content before or after the spoken sentence (e.g., facial expressions, laughter, sync signals, gestures such as covering the mouth) due to non-trimmed segments. However, the core utterance corresponding to the transcript is verified to be uninterrupted.
- In some cases, recordings initially intended for the shared subset were reassigned to the personalized subset if the speaker mispronounced a word, even slightly (e.g., missing articles, incorrect number or pronouns). If the resulting utterance still contains valid Serbian or English words, the transcript was updated accordingly.
- Certain utterances may be semantically odd due to transcript adjustment based on actual speech. Only recordings with clearly invalid words or major issues were removed.
- Some speakers may lack recordings from one of the auxiliary cameras (*video_l* or *video_r*), either entirely or partially.
- Audio recordings contain a tonal component that can be removed via basic signal processing.
- All audio files are mono, 22.05 kHz, in WAV format (PCM_S16LE).
- All video recordings are in MP4 (MPEG-4) format. The frontal camera videos are recorded at 100 fps, while the auxiliary cameras record at 30 fps.
- All participants provided written informed consent for the recording and public release of this dataset, with the option to withdraw their recordings from the public version at any time.