ALOFON
æləfəʊn

ALOFON - Methodology and technology of polymodal allophonic speech transcription

The project aimed to carry out research forming the basis for the development of methods of automatic speech phonetic transcription (in English), based on the use of a combination of information derived from the analysis of audio and video signals. In particular, basic research was conducted on the relationship between allophonic differentiation in speech, i.e., differences in the nature of the same sounds resulting from the different arrangement of articulating organs depending on the phonetic environment (i.e. neighboring phones or prosodic features) and objective signal parameters. In the focus of the researchers' attention were also the parameters of the speech signal (acoustic and vision) characteristic of Poles learning English, including those acquired using the mouth movement system, which allows obtaining additional data to deepen the analysis relating to how to pronounce sounds. The assumption was to develop such detailed analysis methods that they would allow to differentiate minor allophonic and accent variations. As a result of the conducted research it was shown that thanks to the combined analysis of video and audio signals, speech phonetic transcription can be carried out with greater accuracy than using only the acoustic modality, as described in previous works, available in the literature. In-depth research on the diversity of sounds in the context of acoustic and visual signals parameters contributed to the advancement of state of the art in the field of audiovisual speech recognition, and thus in the field of human-computer interaction. In the course of the research, the following hypotheses were experimentally verified: 1. The combined analysis of acoustic and vision data improves the efficiency of phonetic transcription of speech at the allophonic level. 2. The method of speech signal analysis allows a deeper analysis of advanced phonetic aspects of speech compared to the previous state of knowledge. 3. Allophonic aspects, such as nazalization, rounding of vowels, aspiration, characteristic features of lateral allophones, can be effectively detected by analyzing video signals. 4. Differences in the speech signal resulting from allophonic and accent variations can be modeled with the use of appropriate mathematical tools, as well as recognized using machine learning methods. The practical result of the project is also the developed database of recordings, available at the address: http://www.modality-corpus.org/. The MODALITY database consists of over 30 hours of multimodal recordings, including stereoscopic video streams of high resolution, and audio signals obtained from the matrix of the microphone and the internal microphone, embedded in a notebook-class computer. The corpus can be used to develop the AVSR (audiovisual speech recognition) systems because each statement was manually annotated. The developed Modality corpus extension, namely Alofon Corpus, contains audiovisual data and face motion capture data. The audio folder contains the audio tracks recorded by both the directional microphone and the Lavalier microphone. The Vicon folder contains the corresponding files for each recording of the camera system capturing facial muscle movements.

The results of the project were disseminated in the form of 12 publications in scientific journals and in 13 conference papers.

Research sponsored by the Polish National Science Centre, Dec. No. 2015/17/B/ST6/01874.