A Framework for Recording Audio-Visual Speech Corpora with a Microphone and a High-Speed Camera

Alexey Karpov[1,2], Irina Kipyatkova[1], and Miloš Železný[3]

[1] St. Petersburg Institute for Informatics and Automation of RAS, St. Petersburg, Russia
[2] ITMO University, St. Petersburg, Russia
[3] University of West Bohemia, Pilsen, Czech Republic
{karpov,kipyatkova}@iias.spb.su, zelezny@kky.zcu.cz

Abstract. In this paper, we present a novel software framework for recording audio-visual speech corpora with a high-speed video camera (JAI Pulnix RMC-6740) and a dynamic microphone (Oktava MK-012) Architecture of the developed software framework for recording audio-visual Russian speech corpus is described. It provides synchronization and fusion of audio and video data captured by the independent sensors. The software automatically detects voice activity in audio signal and stores only speech fragments discarding non-informative signals. It takes into account and processes natural asynchrony of audio-visual speech modalities as well.

Keywords: Audio-visual speech recognition, multimodal system, automatic speech recognition, computer vision, high-speed video camera.

1 Introduction

At present, there are state-of-the-art studies in audio-visual speech recognition (AVSR) for a lot of languages of the world including English, French, German, Japanese, Chinese, etc. Development of automatic Russian speech recognition technologies based on audio information obtained from a microphone (or a smartphone) is carried out in some International industrial companies such as Google Inc. (Google Voice Search), Nuance (Dragon Naturally Speaking), IBM, in some Russian companies including Speech Technology Center, Stel, Auditech, IstraSoft, Poisk-IT, Cognitive Technologies, Art Brain, SpeechDrive, Kvant, Speech Communication Systems, Speereo, Sistema-Sarov, as well as in several institutes of the Russian Academy of Sciences (SPII RAS, IPU RAS), and leading Universities (SPbSU, MSU, MSLU, ITMO Universities in Russia, Karlsruhe Institute of Technology and University of Ulm in Germany, LIMSI-CNRS in France, University of Aizu in Japan, Binghamton University in USA, etc.), and some other organizations [1-4]. There exists a consortium "Russian Speech Technology" (http://speechtech.ru) that connects main developers of Russian speech recognition systems, but in last years it has been inactive because of some organizational problems.

At that it is well known that audio and visual signals of speech supplement each other very well and their joint multimodal processing can improve both accuracy and

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 50–57, 2014.

robustness of automatic speech recognition (ASR) [5-6]. There are a few of studies of visual-based speech recognition (automatic lip-reading) for Russian in Moscow State University [7], Linguistic University of Nizhny Novgorod [8], in the Higher School of Economics [9], as well as in the Institute of Cybernetics in Ukraine [10]. But at present only in SPIIRAS there are ongoing systematic studies on fusion of audio and video speech modalities for the task of Russian speech recognition [11-14]. We should notice also that recently "RealSpeaker Lab." company (resident of Skolkovo, http://www.realspeaker.net) was also founded. Software called RealSpeaker for video processing was integrated with Google Voice Search ASR and now it supports several languages of the world including Russian. However, there are no scientific papers or technical documentation (besides of advertising materials) with the description of the system that allows us to doubt that any real fusion of audio and video information in the speech recognizer is made. In this system, the free available Internet service Google Voice Search is applied for automatic speech recognition (ASR) and video processing is carried out by a software (based on the free OpenCV library) as a distraction without real integration with audio recognition results and without improving the speech recognition performance.

Stochastic modeling of acoustic and visual speech units is very important for statistical-based methods of ASR. For this purpose, speech databases (corpora) recorded in the conditions approached to the real field conditions as much as possible are needed. For training of the unimodal Russian speech recognition systems a number of speech corpora (in particular, RuSpeech, SPEECHDAT, ISABASE, SPEECON and even database of children speech InfantRU/ChildRU [1]) are already created and commercially available. Recently, a corpus of audio-visual Russian speech (RusAVSpeech-Corpus - Multimedia corpus of audio-visual Russian speech; it was registered in Rospatent in 2011 with № 2011620085) was recorded using a standard video camera (25 frames per second - fps). There is also a multimedia Russian speech corpus called MURCO [15], which is a part of the Russian National Corpus (www.ruscorpora.ru), but it is aimed for studies of emotional and gestural cues in conversational speech and does not contain phonemic and visemic levels of segmentation required for speech recognizer training.

It is important also that recently there were a few of papers reporting on results of automatic lip reading with application of a high-speed video camera [16, 17]. The given researches are focused only on the analysis of visual speech (lip articulations) without audio signal processing. These papers report also on improvement of visual speech recognition accuracy with fps rate higher than 30Hz. There is also a multimodal corpus of Dutch speech recorded by AVT Pike F032C camera with fps above 100Hz [18].

In the end of 2012, an idea was proposed to apply high-speed video cameras along with microphones for the task of audio-visual speech recognition [19]. High frequency of video frames is crucial for analysis of dynamical images, since visible articulation organs (lips, teeth, tip of tongue) change their configuration quite fast at speech production and duration of some phonemes (e.g. explosive consonants) is within 10-20ms (duration of each video at 25 fps frame is 40 ms that is too long). So recordings made by a standard camera with 25-30 fps cannot catch fast dynamics of lips movements and most of the important information is missing in these signals.

In the given paper, we present a novel software framework for recording audio-visual speech corpora with the use of JAI high-speed camera with 200 fps and a dynamic microphone.

2 Framework for Recording Audio-Visual Speech Corpora

The new software framework has been developed in order to capture and record audio-visual speech. One Oktava MK-012 microphone and one JAI Pulnix RMC-6740GE high-speed camera are used. JAI RMC-6740GE is a high-speed progressive scan CCD (charge-coupled) camera. The frame rate for a full resolution image (640x480 pixels) is 200 fps. While recording speaker's face, the camera is rotated by 90 degrees in order to have 480 pixels in horizontal and 640 pixels in vertical that better fits for human face proportions, recorded images are rotated later by the software. This camera transmits all captured raw video data via the Gigabit Ethernet interface. The main problem connected with it consists in data traffic, we use 24-bit color images with the resolution of 640x480 pixels (one uncompressed image takes 0.92MB) in 200 fps mode, and a state-of-the-art PC is not able to save these video data in the real-time mode to HDD and it skips some frames. Nevertheless, it can save whole video data without loss to RAM memory, which has limits on a regular PC and can store only 1-2 min of video data (1 min of video in such a format takes about 11 Gb RAM). It is why we have developed own software to solve this problem.

Architecture of the software framework is presented in Figure 1. The key module is an audio-based voice activity detector (VAD). Audio capturing thread gets signal from the microphone by frames (5ms long) and stores them in a memory buffer, then it checks whether it is speech or not calculating segment energy and applying logical-temporal processing (human being's speech must be longer than 20 consecutive segments and pauses for explosive consonants within a speech fragment must be shorter than 10 segments). If the VAD detects a speech signal it then sends a command to make a time stamp of speech beginning for further audio-visual signal saving.

At the same time and in parallel to this process, video frames captured from the camera are stored in a circular buffer (it is allocated in advance to store 20 sec of last continuous video data) so in any moment it has last 20 sec of visual speech. When the VAD detect end of speech-like signal (i.e. number of pause segments is quite high) then it gives a command to calculate a time stamp of speech ending. It must be noted that continuous speech phrase has to be shorter than 20 sec in other case circular buffer does not contain speech beginning. Then raw video and audio data, which are stored in the corresponding memory buffers, are flushed to the hard disk drive and saved in files. It takes time a bit more than real-time (duration of audio-visual speech). When the software flushes the data then speaker should be mute. On completion saving video and audio data, the speaker can continue the recording session and a next text phrase to read is presented to him/her on the monitor screen, which is located just below the video camera. Here the recording cycle starts from the beginning again and the VAD module waits for a speech signal. When the current speaker has pronounced all prepared phrases, the software stops audio and video capturing threads and keeps working in the off-line mode processing audio and video data stored in the files of the database.

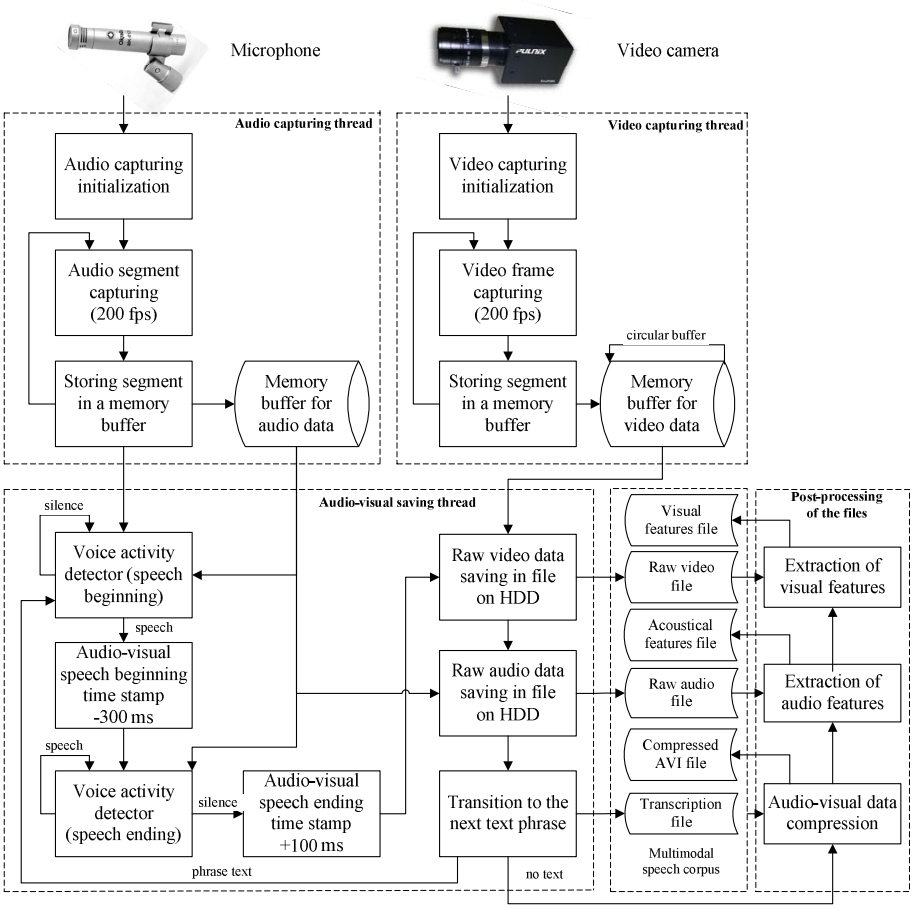

Fig. 1. Architecture of the software framework for recording audio-visual speech corpora

In our audio-visual speech corpus, each speaker has to read and pronounce 100 continuous phrases including 50 phonetically rich and meaningful phrases of the Russian State Standard No. 16600-72 “Speech transmission in radiotelephone communication channels. Requirements for speech intelligibility and articulation evaluation” consisting of 4-8 words plus 50 strings of connected digits (sequences of phone numbers) consisting of 3-7 digits. The former part of the corpus is needed for AVSR system training and the latter one is intended for its adjustment and recognition performance evaluation. An example of recorded audio-visual speech data is provided (http://www.spiiras.nw.ru/speech/JAIavi.zip).

All the recording sessions are made in a quiet office room at daylight, at that all the luminescent lamps must be turned off because they can make artifacts in visual data. Every recording session lasts 30-35 min in average that allows us to get 8-10 min of speech data from each speaker excluding pauses, which are discarded by the VAD module and time for raw data flushing to files and signal post-processing. At first the

recorded audio and video are stored separately in camera's raw format without a compression. At the post-processing on completion of the main recording session the software fuses both files related to one spoken phrase in one avi file with a compression. Also parametrical features are extracted from uncompressed audio and video.

3 Automatic Processing of Audio-Visual Speech Data

The audio signal is captured in mono format with 48kHz sampling rate, SNR ≈ 30dB using Oktava microphone located at 15-20cm from the speaker's mouth. As acoustical features, 12-dimentional Mel-Frequency Cepstral Coefficients (MFCC) are calculated from 26 channel filter bank analysis of 20ms long frames with 5ms step [20]. Thus, the frequency of audio feature vectors is 200Hz.

Visual parameters are calculated as a result of the following signal processing steps using computer vision library OpenCV [21]: multi-scale face detection in video frames with 200 fps using a cascade classifier with AdaBoost method based on the Viola-Jones algorithm [22] with a face model; mouth region detection with two cascade classifiers (for mouth and mouth-with-beard) within the lower part of the face [23]; normalization of detected mouth image region to 32×32 pixels; mapping to a 32-dimentional feature vector using the principal component analysis (PCA); visual feature mean normalization; concatenation of the consecutive feature vectors into one vector to store the dynamic information in the feature data; viseme-based linear discriminant analysis (LDA). The video signal processing module produces 10-dimentional articulatory feature vectors with 200Hz frequency.

An important issue at audio-visual speech synchronization and fusion in any AVSR system is a natural asynchrony of both speech modalities. Audible speech units (phones) and visible ones (visemes) are not completely synchronized in human speech. It is partially caused by the human's speech production system. Inertance of the vocal tract and its articulation organs results in the co-articulation phenomenon, which reveals itself differently on two speech modalities and causes some asynchrony between them. Recent corpus-based studies reported [14, 24], that that visemes always lead in phone-viseme pairs as well as, at a beginning of a phrase visual speech units usually lead more noticeably (up to 150-200 ms for stressed rounded vowels) over the corresponding phonemes than in the central or ending part of the phrase.

Some studies also showed [25, 26], that degree of asynchrony of phoneme and viseme stream in the speech formation process is different for different languages and nations. For example, for Japanese lip movements and sound stream of speech are almost synchronous, therefore early method of audio-visual speech fusion shows the best results. In English (especially American English) there is rich lip articulation (even hyper-articulation) that often causes temporal disagreement between streams of the formed phonemes and visemes.

In order to take into account this phenomenon for recording audio-visual speech corpus, the software makes a time stamp of speech beginning 300 ms before the first speech segment detected by the VAD, so audio-visual speech fragment longer in 300 ms is taken in order to catch beginning of lips movements involved in audible speech production. The same audio-visual speech capturing/recording software framework

can be also used for VAD and signal capturing in the on-line speech recognition mode with direct use of microphone and camera.

In order to automatically generate initial phoneme and viseme segmentations of the data, we apply hidden Markov models (HTK toolkit) with forced alignment method based on the Viterbi algorithm [20]. This method uses canonic text transcriptions of the spoken phrases made by an expert and feature vectors calculated by the system. As a result of the forced alignment, audio-visual speech signals are matched with corresponding transcriptions and optimal signal segmentations with time stamps for the audio and visual speech units are produced. After that automatically made segmentations have to be checked and corrected by an expert.

At present, we are in the process of recording our audio-visual Russian speech corpus and we have already collected of speech data from several Russian speakers having normal voice and articulation, both men and women. Multimodal data (video, audio, texts) of each speaker takes about 50–80GB, 99.8% of which are video data. After this stage we will develop an AVSR system for Russian, which can apply both the microphone and the high-speed video camera. General architecture of the AVSR system is presented in Figure 2. It will use state asynchronous Coupled Hidden Markov Models (CHMM) for stochastic modeling audio and video signals of speech [13]. Later this system will be a part of a universal assistive information technology [27].

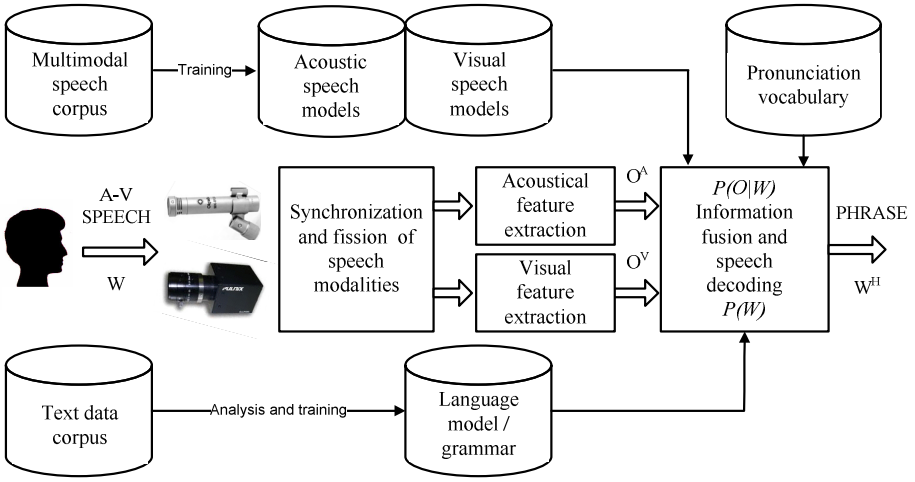

Fig. 2. General architecture of the audio-visual Russian speech recognition system (AVRSRS)

4 Conclusion and Future Work

In the paper, we have described the software framework for audio-visual speech signal recording and processing. This software helps at recording audio-visual speech corpora and uses JAI Pulnix RMC-6740 high-speed video camera (it allows capture video frames with 200 fps) and Oktava MK-012 microphone. The software

framework provides synchronization and fusion of audio and video data captured by the independent microphone and video camera, it automatically detects voice activity in audio signal and stores only speech fragments discarding non-informative signals, as well as it takes into account and processes natural asynchrony of audio-visual speech modalities as well. Our software can be also used for voice activity detection and signal capturing in the on-line mode of AVSR with direct use of the audio-visual sensors. At present, we are in the process of recording our audio-visual Russian speech corpus and we have already collected of speech data from several Russian speakers. Multimodal data (video, audio, texts) of each speaker takes about 50–80GB, 99.8% of which are video data. After this stage we will develop an AVSR system for Russian that can apply both the microphone and the high-speed camera and will use state asynchronous CHMM for modeling audio and video signals.

Acknowledgments. This research is partially supported by the Russian Foundation for Basic Research (Project № 12-08-01265-a), by the Government of Russian Federation (Grant 074-U01), as well as by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090.

References

1. Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A.: Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Communication 56, 213–228 (2014)
2. Kipyatkova, I., Verkhodanova, V., Karpov, A.: Rescoring N-Best Lists for Russian Speech Recognition using Factored Language Models. In: Proc. 4th International Workshop on Spoken Language Technologies for Under-resourced Languages SLTU-2014, St. Petersburg, Russia, pp. 81–86 (2014)
3. Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M.: Modeling of Pronunciation, Language and Nonverbal Units at Conversational Russian Speech Recognition. International Journal of Computer Science and Applications 10(1), 11–30 (2013)
4. Kipyatkova, I., Karpov, A.: Lexicon Size and Language Model Order Optimization for Russian LVCSR. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS (LNAI), vol. 8113, pp. 219–226. Springer, Heidelberg (2013)
5. Potamianos, G., et al.: Audio-Visual Automatic Speech Recognition: An Overview. Chapter in Issues in Visual and Audio-Visual Speech Processing. MIT Press (2005)
6. Bailly, G., Perrier, P., Vatikiotis-Bateson, E.: Audiovisual Speech Processing. Cambridge University Press (2012)
7. Soldatov, S.: Lip reading: Preparing feature vectors. In: Proc. International Conference Graphicon 2003, Moscow, Russia, pp. 254–256 (2003)
8. Gubochkin, I.: A system for tracking lip contour of a speaker. In: Modern Science: Actual problems of theory and practice. Natural and Technical Sciences, No. 4-5, pp. 20–26 (2012) (in Rus.)
9. Savchenko, A., Khokhlova, Y.: About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems. Optical Memory and Neural Networks (Information Optics) 23(1), 34–42 (2014)

10. Krak, Y., Barmak, A., Ternov, A.: Information technology for automatic lip reading of Ukrainian speech. Computational Mathmatics. Kyiv 1, 86–95 (2009) (in Rus.)
11. Železný, M., Císar, P., Krnoul, Z., Ronzhin, A., Li, I., Karpov, A.: Design of Russian audio-visual speech corpus for bimodal speech recognition. In: Proc. 10th International Conference on Speech and Computer SPECOM 2005, Patras, Greece, pp. 397–400 (2005)
12. Cisar, P., Zelinka, J., Zelezny, M., Karpov, A., Ronzhin, A.: Audio-visual speech recognition for Slavonic languages (Czech and Russian). In: Proc. International Conference SPECOM 2006, St. Petersburg, Russia, pp. 493–498 (2006)
13. Karpov, A., Ronzhin, A., Markov, K., Zelezny, M.: Viseme-dependent weight optimization for CHMM-based audio-visual speech recognition. In: Proc. Interspeech 2010 International Conference, Makuhari, Japan, pp. 2678–2681 (2010)
14. Karpov, A., Ronzhin, A., Kipyatkova, I., Zelezny, M.: Influence of phone-viseme temporal correlations on audio-visual STT and TTS performance. In: Proc. 17th International Congress of Phonetic Sciences ICPhS 2011, Hong Kong, China, pp. 1030–1033 (2011)
15. Grishina, E.: Multimodal Russian corpus (MURCO): First steps. In: Proc. 7th Int. Conf. on Language Resources and Evaluation LREC 2010, Valetta, Malta, pp. 2953–2960 (2010)
16. Chitu, A.G., Rothkrantz, L.J.M.: The influence of video sampling rate on lipreading performance. In: Proc. SPECOM 2007, Moscow, Russia, pp. 678–684 (2007)
17. Chitu, A.G., Driel, K., Rothkrantz, L.J.M.: Automatic lip reading in the Dutch language using active appearance models on high speed recordings. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 259–266. Springer, Heidelberg (2010)
18. Chitu, A.G., Rothkrantz, L.J.M.: Dutch multimodal corpus for speech recognition. In: Proc. LREC 2008 Workshop on Multimodal Corpora, Marrakech, Morocco, pp. 56–59 (2008)
19. Karpov, A., Ronzhin, A., Kipyatkova, I.: Designing a Multimodal Corpus of Audio-Visual Speech using a High-Speed Camera. In: Proc. 11th IEEE International Conference on Signal Processing ICSP 2012, pp. 519–522. IEEE Press, Beijing (2012)
20. Young, S., et al.: The HTK Book, Version 3.4. Cambridge Univ. Press (2009)
21. Liang, L., Liu, X., Zhao, Y., Pi, X., Nefian, A.: Speaker independent audio-visual continuous speech recognition. In: Proc. Int. Conf. on Multimedia & Expo ICME 2002, Lausanne, Switzerland, pp. 25–28 (2002)
22. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition CVPR 2001, USA, pp. 511–518 (2001)
23. Castrillyn, M., Deniz, O., Hernndez, D., Lorenzo, J.: A comparison of face and facial feature detectors based on the Viola-Jones general object detection framework. Machine Vision and Applications 22(3), 481–494 (2011)
24. Feldhoffer, G., Bardi, T., Takacs, G., Tihanyi, A.: Temporal asymmetry in relations of acoustic and visual features of speech. In: Proc 15th European Signal Processing Conference EUSIPCO 2007, Poznan, Poland, pp. 2341–2345 (2007)
25. Sekiyama, K.: Differences in auditory-visual speech perception between Japanese and America: McGurk effect as a function of incompatibility. Journal of the Acoustical Society of Japan 15, 143–158 (1994)
26. Chen, Y., Hazan, V.: Language effects on the degree of visual influence in audiovisual speech perception. In: Proc. 16th International Congress of Phonetic Sciences ICPhS 2007, Saarbrücken, Germany, pp. 2177–2180 (2007)
27. Karpov, A., Ronzhin, A.: A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2014, Part I. LNCS, vol. 8513, pp. 369–378. Springer, Heidelberg (2014)