Audio-Visual Speech Recognition for Slavonic Languages (Czech and Russian)

Petr Císař[1], Jan Zelinka[1], Miloš Železný[1], Alexey Karpov[2], Andrey Ronzhin[2]

[1]Department of Cybernetics, University of West Bohemia in Pilsen (UWB), Czech Republic
{pcisar1,zelinka,zelezny}@kky.zcu.cz

[2]Speech Informatics Group, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Russian Federation
{karpov,ronzhin}@iias.spb.su

Abstract

The paper presents the results of recent experiments with audio-visual speech recognition for two popular Slavonic languages: Russian and Czech. The description of test applied tasks, the process of multimodal databases collection and data pre-processing, methods for visual features extraction (geometric shape-based features; DCT and PCA pixel-based visual parameterization) as well as models of audio-visual recognition (concatenation of feature vectors and multi-stream models) are described. The prototypes of applied systems which will use the audio-visual speech recognition engine are mainly directed to the market of intellectual applications such as inquiry machines, video conference communications, moving objects control in noisy environments, etc.

1. Introduction

At present the research and development of the systems for speech recognition are carried out mainly for English, Chinese and Japanese. Also there are successful enough researches on speech recognition for several Romanic languages (for French and German). However there are not available speech recognition systems for Slavonic group of European languages.

But the most part of Eastern Europe and Central Europe speaks on Slavonic languages. Inside Slavonic linguistic family three groups can be selected: (1) East-Slavonic group: Russian (number of speakers is over 100 millions of citizens in the European part of Russia), Ukrainian (40 millions), Byelorussian (9 millions); (2) West-Slavonic group: Polish (27 millions), Czech and Slovak (15 millions); (3) South-Slavonic group: Bulgarian (7 millions), Serbo-Croatian (about 10 millions), Slovenian (about 1,5 millions). Grand total over 210 millions of citizens of Europe (over 25%) use Slavonic languages for daily human-human communications, but real systems for human-computer interaction in Slavonic languages are demonstrated very rarely. This research aims to fill the gap in introducing and promoting advanced speech technologies for main Slavonic languages (mainly, for Russian and Czech).

All the Slavonic languages are very similar in lexical structure and mechanisms of word-formation. All the Slavonic languages belong to the category of synthetical languages. Synthetical languages are characterized by the tendency to combination (synthesizing) of the lexical morpheme (or several lexical morphemes) and one or several grammatical morphemes in one word-form. To develop an effective speech recognition system for any language it is required to evaluate the peculiarities of these languages in comparison with other groups of languages and analyze the problems appeared during the development of a speech recognition system.

This paper presents an advanced kind of speech recognition techniques. It is the recognition of speech based on joint information from audio waves of an utterance and visual information of lips movements that is produced during uttering any word by a human. The importance and effectiveness of combining the audio and visual modalities is clear. These kinds of information supplement each other in different conditions and increase performance of audio-visual speech recognition (AVSR) system: (1) In noisy environment the audio information is not enough and visual information allows increasing the speech recognition accuracy. In conditions with low level of audio-signal and noisy environment or unwanted sounds the standard speech recognition systems cannot provide the required accuracy. In order to increase robustness of speech recognition it is necessary to use the visual information in addition to audio signal. (2) In poor illumination environment the audio information can fill the gap of visual information.

The ideas of creation of AVSR system for Slavonic languages and the process of databases collection were presented at previous SPECOM conference [1]. This paper gives the results of recent experiments on audio-visual recognition for Russian and Czech languages.

The development of AVSR was made jointly by UWB and SPIIRAS in framework of the SIMILAR Network of Excellence as scientific exchanges between organizations during 2005-2006.

2. Prototypes of audio-visual speech recognition systems

2.1. Czech speech recognition task

The main reason for using of visual speech information is to improve the system of audio speech recognition in noisy environments. One of the applications of this system can be recognition of car driver utterances. The car produces sufficient noise for decreasing of recognition rate of audio speech recognition. We collected audio-visual database of car driver utterances for testing of audio-visual recognition. We developed the system of liptracking for this database [2]. The system works well but we found out that the database is not suitable for testing of visual parameterization. The visual

parameterization is the most important thing in the process of speech recognition. We had to collect new audio-visual database to develop new visual parameterization. The database was made in laboratory conditions. We would like to create new visual parameterization method which is based on expert's knowledge. We suppose that shape-based visual parameterization can be better than pixel-based visual parameterization. If we want to compare our new method and common used pixel-based parameterization we have to make the test on XM2VTS database. There are some results of AVSR experiments with pixel-based visual parameterization on this database (section 6.1).

The design of UWB-04-HSCAVC (UWB stands for university of West Bohemia, 04 for year of recording, HSCAVC for Hundred Speakers Czech Audio-Visual Corpus) corpus was based on the experiments gained as a result of recording and experimenting with previous corpora, mainly UWB-03-CIVAVC corpus [2]. Also, we studied published papers on this topic. The database presented in [3] seems to be well designed and even though it is designed for different purpose, it served as reference corpus during our design. The database is primarily intended for research on visual speech parameterizations. For this reason we retained optimal recording conditions during the whole recording, such as constant illumination, static head position, or front view. The corpus consists of recordings of 100 speakers, 39 males and 61 females. The average age of speakers is 22 years. Recording of one speaker was done during one session. Speaker was asked to read sentences from auxiliary screen. The average total length of recording for one speaker was 23 minutes, which makes together more than 38 hours of speech.

The text of a corpus consists of 200 sentences for each speaker. First 50 sentences are common for all speakers. They are selected to contain all phonemes. The rest of 150 sentences are different for each speaker. They are balanced to contain as many triphones as possible.

The corpus is not divided into training and testing sections. We suppose that 20 speakers out of the total 100 will be taken out from the whole database and used as testing speakers while the rest 80 will be used as the training ones. The gender distribution should be retained in both parts. Two or more different setups (divisions to the training and testing part) can be used during experiments. The audio-visual data in the corpus are supplemented with the annotations and preprocessing data. The first part consists of annotations. All audiovisual data are transcribed into transcription files (.trs format). These data contain information about the start and the end of each sentence. Using this information the audio-visual data are split to single sentences.

2.2. Russian speech recognition task

The AVSR system can be successfully used for development of intellectual applications such as smart phones, tourist information systems, car equipment control, video conference communications etc. The most perspective market for commercial use of such systems is automated inquiry systems, which are located in hotels, airports, train stations, subway, supermarket, tourist places. Such machines-kiosks allow getting the information on local services, ground transportation, parking, shopping, entertainment. Using the AVSR we can improve traditional touch-screens by voice access, effective in any conditions (acoustic noise, darkness, etc).

As test task for speech recognition the task of voice access to the electronic catalogue "Yellow Pages of Saint-Petersburg" (see "http://www.yell.ru") was used. This catalogue contains the list of all organizations with reference to address, phone number and kind of activity (rubric). The total number of rubrics in this service is above 1600. As the final result of research the united automatic information service, which will help find an address or phone number of necessary organization, will be developed. In this system the audio speech recognition and visual recognition modules are used for processing a user's requests. A user will be able to realize two types of requests to the automatic system: (1) search of an organization by rubric name (kind of activity). In this mode a user may say a name of an interesting topic (for example, "Restaurants of Japanese cuisine") and the system finds and generates the list of organizations corresponding to this topic. (2) search of an organization by its attributive information: company's name, address or phone number. In this mode a user should choose the type of search before and then pronounce the available information. For audio speech recognition with large vocabulary the SIRIUS ASR engine is being developed by SPIIRAS [4].

To test the models of AVSR system it is not required to use large vocabulary and therefore 50 main rubric names, containing 102 diverse Russian words, were selected for tests.

Multimodal audio-visual database was collected for Russian language. Ten male speakers were recorded in office conditions. The average age of speakers is 20 years. Each speaker has to read 200 phrases (names of the rubrics of the catalogue "Yellow Pages of Saint-Petersburg") with the length from 1 till 5 words. Recording time for a speaker is approximately 15 minutes. This database was divided into two parts: 80 percents of utterances of each speaker were used for training purpose and other rest of data for model testing.

3. Image pre-processing stage

The image preprocessing data consist of the static data and dynamic data. First, for all speakers the face is found. Then the region containing only skin colored parts of the face is extracted and stored. The mean value of the skin color of the face is stored separately in a text file. Also, the regions for eyes and nose are extracted and stored in separate image files.

Then each video file is processed as a whole. The head is found using firstly converting the image into CR/CB space and then thresholding. The face region is detected as a convex hull of the thresholding result. At this moment we know the position and roughly the orientation of a head. We can find the region of the eyes and the mouth. The algorithm detecting the position of lips is similar to the one used in [5].

The algorithm continues as follows. In first frame the eyes are found and their distance is measured. The size of the region of interest of lips is then set according to this distance. For each subsequent frame then the center of lips and eyes are measured and used for detecting the region of interest, as shown in Figure 1. This way we obtain size and rotation invariant region of interest. During the detection of eyes we check whether the eyes are not closed. In case of closed eyes, we do not detect their position in that frame and rather use their position from previous frame. For each video file we get the file with its description containing information about

mouth center, its rotation and the size of region of interest. Using this information the region of interest can be easily extracted from each video frame during processing. The design of an experiment thus can focus on parameterization.

This preprocessing stage was applied for Czech and Russian speech databases as well as to the XM2VTS database which is used in experiments with visual parameters.

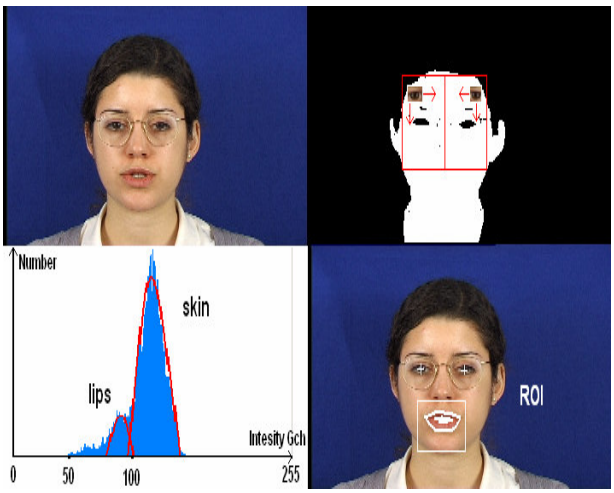


Figure 1. Localization of head, eyes and lips

At the first step of research the multimodal databases containing the audio and visual parts were collected. Based on these data the models of AVSR are trained and tested with diverse types of visual features and methods of early fusion of the acoustic features (MFCC features) and the visual features of the region of mouth, which are described below.

4. Visual feature extraction

4.1. Geometric shape-based parameterization

Our geometric parameterization is based on a description of the shape of lips and the positions of tongue and teeth. Therefore we have to detect outer and inner lip contours for each frame.

The process of lip-tracking should be fast because we have to process as many frames per second as possible. We decided to use a simple and fast thresholding method to get a binary image of lips. We work just with the ROI. The position of ROI was estimated during pre-processing of the corpus. The ROI is represented in chromatic colors. We use just the part Gch (green). R, G, B are parts of RBG color space.

$$G_{ch} = \frac{G}{R+B+G} \tag{1}$$

We use chromatic color to avoid the influence of illumination. The threshold is estimated by an automatic clustering method based on GMM algorithm. We tried to get the threshold by the analysis of histogram [2] but the method worked only for good conditions. In the ROI there are two main objects (mouth and skin). Therefore the clustering algorithm decides the pixels of ROI to two clusters and gives as a variance and a mean value of these objects. We can calculate the threshold as:

$$t = m_L + \left(\frac{(m_S - m_L)}{n_S + n_L} * n_1 \right), \tag{2}$$

where mL is the mean intensity value of lips, mS is the mean intensity value of skin, nS is the number of points of lips, nL is the number of points of skin.

The clustering algorithm is relatively time consuming. Therefore we calculate the threshold for the first frame only and use it for whole sentence. When this threshold is applied we get the binary image of lips as depicted in Figure 2a. The object corresponds with the outer lip contour well, but sometimes does not correspond with the inner lip contour well. The crucial problem connected with thresholding of tongue and dark places inside the lips. We process the inner part of lips one more time. We focused on the area of inner lips during second processing. This area is represented by the outer boundary of lips. We know the vertical center of lips. The algorithm deals with line which connects upper and lower centers of outer lips boundary. We look for local minima which represents inner lips boundary on this line. If we have the value of intensity of this inner boundary we determine the threshold for inside of mouth. This threshold is used to get the binary image of inner part of mouth. The final shape of inner part is repaired by this binary image as depicted in Figure 2d. Now we have the object of lips, but there are many errors in the shape. Some parts are missing and some are redundant.

It is necessary to repair whole shape of lips. We use an active shape model of lips which consist of 32 points and it is controlled by PCA parameters. The model can be deformed to every possible shapes of lips.

If we want to deform the model to possible shapes of lips we have to establish training set which contains examples of these shapes. It was chosen 200 templates of lips for 30 different speakers.

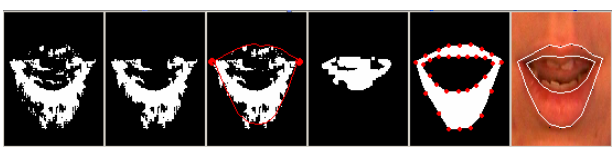
Inner and outer contour was marked for each template and then 16 points of inner and 16 points of outer lips boundary were found automatically. The model is normalized on the width of outer lips. We can do it because the corners of lips are localized very well in each frame. The points of inner contour are normalized by the width of outer lips then these points are centered by the center of outer lips corners. We calculated 10 PCA coefficients and transformation matrix in the same way as in [6]. Now it is possible to deform the model by these 10 control coefficients and the transformation matrices.

We use the binary image of lips. We localized 32 points of the model on this binary image as depicted in Figure 2e. Then we transform the model to the lower dimension and back to 32 points by transformation matrices. The shape of the model is smoothed and repaired as depicted in Figure 2f. There is another problem during determination of the binary object of lips. There is more than one object after thresholding of the lips. The object of lips should be the biggest object in the picture but sometimes the lips are separated to many objects as depicted in Figure 2a. When the biggest object is selected then we can lost some parts of upper lips as depicted in Figure 2b. We use the model of lips from previous frame to determine which parts belong to the lips and which to the background as depicted in Figure 2c. This step usually repairs the upper part of lips.

The first part of parameterization (geometric) is calculated from determined contour of the lips. We chose 4 features for

description of absolute size of the lips. That is height and width of outer and inner contour of the lips. We chose these four parameters because it is easy to normalize them over whole sentence to eliminate the dependence on the speaker. Next 10 features describe the shape of the lips independently on the size of the speaker's mouth. These features are 10 PCA control parameters of the model.

The next part of the parameterization is the description of inner part of the mouth. We know that the visibility and relation between upper teeth, lower teeth and tongue are very important for lip-reading. It was not possible to use our algorithm for localization of the part of inner mouth because of the resolution of the ROI. Therefore we had to use the DCT parameterization for the description of inner mouth. We chose the part of the mouth as training data for determination of the 4 DCT features which should describe four objects inside the mouth. The part of the mouth is limited by the outer contour of lips depicted in Figure 2. Our final parameterization (hybrid visual features) consists of 18 visual parameters.



(a) (b) (c) (d) (e) (f)

Figure 2. Localization of the lips contour: a) thresholded image, b) biggest binary object, c) repairing of the binary object, d) inner mouth, e) localization of the model points, f) repaired shape of lips

4.2. DCT pixel-based parameterization

Pixel based parameterization was calculated from the ROI. We know the position and size of ROI from pre-processing of the database. We used two thousands of ROI for determination of 10 DCT features. We calculated 64 DCT parameters for all the training images. The DCT features were selected as the fist 10 DCT parameters with the highest energy. Than we calculated the DCT features for testing data and this set is final DCT parameterization like in [7].

4.3. PCA pixel-based parameterization

This kind of pixel-based features is realized in Intel OpenCV Open Source Computer Vision Library [8]. It was tested for AVSR also. In this parameterization method the obtained images of mouth region (ROI) are normalized to 32×32 in size the gray level pixels in the mouth region are mapped into a 32-dimensional feature vector using the principal component analysis (PCA). The PCA projection is computed from a set of two thousands of training mouth region images from the database.

For calculation of PCA projection the following data are used: $U = (u^1, u^2, ..., u^N)^T$ is a N-dimensional vector containing the pixels of an image with the size $W \times H$. Having a set of M input vectors $\{U_1, U_2, ..., U_m\}$, the mean vector μ and the covariance matrix C are defined as:

$$\mu = \frac{1}{M}\sum_{k=1}^{m} U^k \tag{3}$$

$$C = \frac{1}{M}\sum_{k=1}^{M}(U_k - \mu)(U_k - \mu)^T = \frac{1}{M}\sum_{k=1}^{M} U_k \cdot U_k^{\,T} - \mu\mu^T \tag{4}$$

Also the sum vector $\tilde{\mu}$ and the partial covariance matrix $\tilde{C}$ are calculated:

$$\tilde{\mu} = \sum_{k=1}^{M} U^k \tag{5}$$

$$\tilde{C} = \sum_{k=1}^{m} U_k U_k^{\,T} \tag{6}$$

The first p largest eigenvalues and the corresponding eigenvectors of PCA are $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_p\}$ and $V = (V_1, V_2, ..., V_p)^T$, the projection of input vector u in the p-dimensional subspace $Y = (y_1, y_2, ..., y_p)^T$ is calculated as:

$$Y = V(U - \mu) \tag{7}$$

The result vector is normalized using eigenvalues:

$$\hat{y}_i = \frac{y_i}{\sqrt{\lambda_i}} \tag{8}$$

The final PCA feature vector is:

$$\hat{Y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_p)^T \tag{9}$$

Each of these visual feature extraction techniques was applied and tested for AVSR and the results of experiments are presented below.

5. Models of audio-visual speech recognition

5.1. Model of concatenation of feature vectors

We used the feature based combination of audio and visual stream. Simple concatenation of the visual and acoustic parameterization was used [9]. An audio signal length and relevant video signal length were equalized using linear interpolation on video signal. Delta and acceleration coefficients were reckoned in both video signal parameterizations. Finally, feature vector mean subtraction and variance normalization was applied per utterance. The both streams were fussed by means of concatenative fusion because we assumed standard modality fusion in multi-stream model.

5.2. Multi-stream model

The multi-stream model of speech recognition was applied for AVSR also. This model belongs to the class of state synchronous decision fusion [10]. The multi-stream model was realized by Hidden Markov Model Toolkit (HTK) [11]. This toolkit allows building and manipulating both single-stream and multi-stream Hidden Markov Models. The main difference between single-stream HMMs, which are used for speech recognition mainly, and multi-stream HMMs consists in diverse calculation of the probability of audio-visual observation vector $o^{(t)}$ in a state c of a multi-stream HMM. This probability can be calculated as follows:

$$P[o^{(t)} \mid c] = \prod_{s \in \{A,V\}} [\sum_{j=1}^{J_{sc}} w_{scj} N(o^{(t)}, m_{scj}, v_{scj})]^{\lambda_{sct}} \tag{10}$$

Here λ_{sct} is the positive stream exponent which depends on the type of the modality *s*, HMM's state *c* and frame of the speech *t*. These modality weights are global and constant over the entire speech database. J_{sc} is the number of mixture components in the stream, w_{scj} is the weight of the *j*-th component and $N(o^{(t)}, m_{scj}, v_{scj})$ is a multivariate Gaussian with mean vector *m* and covariance matrix *v* that equals:

$$N(o^{(t)}, m_{scj}, v_{scj}) = \frac{1}{\sqrt{(2\pi)^n \mid v \mid}} e^{-\frac{1}{2}(O-m)^T v^{-1}(O-m)} \quad (11)$$

where *n* is the dimensionality of the feature vector *O*.

During the training process the modality weights are tuned manually by minimizing the WER of audio-visual recognition model. All other parameters of HMMs are re-estimated by Baum-Welch procedure.

6. Experimental results

6.1. Experiments with visual and audio-visual recognition

We compared DCT parameterization and hybrid visual features on XM2VTSDB database. After exclusion of incompetent or incorrect records from the database, we reserved 1847 records as the training data and 390 recordings as the testing data. Each record in the corpus includes utterance of a sequence of 20 numbers between zero and nine. Moreover, each number in this interval is in the sequence exactly two times.

First we realized visual speech recognition only and then AVSR with concatenation of feature vectors (see Section 5.1). We used the HMM based multi-modal speech recognition system with 0-gram language model, i.e. the constant probability of a word, and synchronous monophone based acoustic models. In our experiments the percent of accuracy and the percent of correctness of speech recognition for the individual words were calculated. Table 1 shows the results of the experiments.

First column contains the results of visual recognition only by DCT parameterization. The second column includes the results obtained using hybrid visual parameterization. The results of speech recognition from noisy audio signal are in the third column. The last column contains the results of the audio-visual recognition. The SNR was about 0 db for both cases. The accuracy for clean audio signal was 98.86 % and the correctness was 99.26 %.

Table 1. Results of audio-visual experiments

	DCT	**hybrid**	**Audio**	**Audio+ hybrid**
Accuracy	48.44 %	55.17 %	91.29 %	94.0 %
Correctness	52.21 %	70.51 %	92.39 %	96.36 %

6.2. Experiments on Russian speech recognition

The multi-stream model (see Section 5.2) was used for audio-visual recognition of Russian speech. For parameterization of audio signal 12 MFCC features were used and geometrical shape-based features (see Section 4.1) as well as pixel-based PCA features were applied for video signal. The Table 2 presents the list of Russian visemes and visemes to phonemes mapping used in audio-visual speech recognizer.

Table 2. The list of Russian visemes used for recognition

Visemes	**Russian phonemes**
s	silence
a	а, а!, е, е!
i	и, и!, ы, ы!
o	о!, у, у!
v	ф, ф', в, в'
z	з, з', с, с', ц, ч
p	м, м', б, п, п'
t	т, т', д, д', н, н', к, к', г, г'
l	л, л', р, р'
j	ж, щ, х, ш, й

The results of experiments (sentences accuracy rate) are presented in the Table 3. The size of vocabulary in this task is 102 words and the table shows the accuracy of word recognition. It can be seen from the table that the AVSR shows the better results than audio only speech recognizer. Meanwhile the results using pixel-based PCA features and shape-based visual features were almost the same. The modality weights were manually adjusted for maximal WER and the weight for video streams was 0.2 and the weight for audio stream is 2.0 (1.8 in second case). The signal-to-noise ratio (SNR) for audio signal was 10 db for all the experiments (clean speech).

Table 3. Results of experiments with Russian AVSR

	Audio	**Audio + pixel-based**	**Audio + shape-based**
Accuracy rate	90.1 %	92.0 %	92.3 %

7. Conclusions

In this article we presented AVSR experiments on Slavonic languages. The audio-visual database UWB-04-HSCAVC for Czech language was collected. We realized basic pre-processing on visual part of the database (determination of region of interest). The algorithm for localization of inner and outer contour of lips was created. We studded the process of lip-reading by human lip readers and we found out that it will be better to describe the visual speech information by shape based parameterization than by pixel based. Therefore we designed our first hybrid visual parameterization. The parameterization consists of geometric description of lips (inner and outer contour) and of DCT coefficients which describe the inner part of the mouth. This hybrid parameterization was used for visual and audio-visual speech recognition experiments on XM2VTS database. We compared the results of hybrid parameterization (video-only 70.51% correctness) with common use DCT parameterization (video-only 52.21%). Hybrid parameterization got better results than DCT parameterization. Now we are going to do the recognition experiments on Czech database. The next step of our research is to use expert knowledge of human lip-reading for designing of a new visual parameterization. This parameterization should describe the process of creation of visemes which is speaker independent. Also the Russian

audio-visual speech recognition system was tested. The experiments using MFCC audio features and shape-based visual features showed the best results on developed Russian audio-visual database.

8. Acknowledgements

This research is supported by the European Community in framework of the SIMILAR European Network of Excellence "The European taskforce creating human-machine interfaces SIMILAR to human-human communication", project No. 507609 as well as by the Grant Agency of the Academy of Sciences of the Czech Republic, project No. 1ET101470416 and by the Ministry of Education of the Czech Republic, project No. MSM235200004.

9. References

[1] 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. of 10-th International Conference "Speech and Computer" SPECOM'2005, Patras, Greece, pp. 397-400.

[2] Císař, P., Železný M., and Krňoul Z., "3D Lip-tracking for Audio-Visual Speech Recognition in Real Applications," In Proc. of the ICSLP 2004, Jeju island, Korea, 2004.

[3] Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., "XM2VTSDB The Extended M2VTS Database," Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA '99), Washington D.C., USA, 1999.

[4] Karpov, A., Ronzhin, A., Li, I. "SIRIUS: A System for Speaker-Independent Recognition of Continuous Russian Speech". In TRTU Proceedings, № 10, 2005, pp. 44-53.

[5] Císař, P., Železný, M., Krňoul, Z., Kanis, J., Zelinka, J., Müller, L. "Design and recording of Czech speech corpus for audio-visual continuous speech recognition", In Proc. of the AVSP 2005. Causal Productions, Adelaide, Australia, 2005.

[6] Cootes, T. F., Taylor, C.J., Cooper, D.H, and Graham, J. "Active Shape Models – Their Training and Application", Computer Vision and Image Understanding, 61(1): 38-59, 1995.

[7] Heckmann, M., Kroschel, K., Savariaux, C., and Berthommier, F., "DCT-based video features for audiovisual speech recognition", In Proc. of the ICSLP 2002, Denver, CO, USA, 2002.

[8] http://www.intel.com/technology/computing/opencv/

[9] Glotin, H., Vergyri, D., Neti, C., Potamianos, G., and Luettin, J. "Weighting schemes for audio-visual fusion in speech recognition", In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'2001, Salt Lake City, USA, 2001, pp. 165–168.

[10] Neti, C., Potamianos, G., Luettin, J., et al., "Audio-visual speech recognition", Final Workshop 2000 Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, 2000.

[11] http://htk.eng.cam.ac.uk/