Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition

Jana Trojanová, Marek Hrúz, Pavel Campr, Miloš Železný

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia
Univerzitni 22, 306 14, Plzen, Czech Republic
{trojana, mhruz, campr, zelezny }@kky.zcu.cz

Abstract

In this paper we discuss the design, acquisition and preprocessing of a Czech audio-visual speech corpus. The corpus is intended for training and testing of existing audio-visual speech recognition system. The name of the database is UWB-07-ICAVR, where ICAVR stands for Impaired Condition Audio Visual speech Recognition. The corpus consist of 10000 utterances of continuous speech obtained from 50 speakers. The total length of the database is 25 hours. Each utterance is stored as a separate sentence. The corpus extends existing databases by covering condition of variable illumination. We acquired 50 speakers, where half of them were men and half of them were women. Recording was done by two cameras and two microphones.

Database introduced in this paper can be used for testing of visual parameterization in audio-visual speech recognition (AVSR). Corpus can be easily split into training and testing part. Each speaker pronounced 200 sentences: first 50 were the same for all, the rest of them were different. Six types of illumination were covered. Session for one speaker can fit on one DVD disk. All files are accompanied by visual labels. Labels specify region of interest (mouth and area around them specified by bounding box). Actual pronunciation of each sentence is transcribed into the text file.

1. Introduction

The automatic speech recognition (ASR) has attracted researchers from early 1960th. From that time it has developed through the recognition of the isolated words from one speaker to speaker-independent continuous speech. Nowadays systems accomplish good recognition rates under limited conditions dependent on the type of the task. The factors that influence the recognition rate most of all are environment and type of the speech. By the environment it is especially meant environmental noise that degrades acoustic signal. In this case the visual part of the speech can be used to increase recognition rate. Such a system is called an audio-visual speech recognition (AVSR). This paper deals with design, recording and preprocessing of the audio-visual speech database.

There are many databases for ASR recognition that contain big number of speakers, different types of utterances and degradations of the signal. In the domain of AVSP only a few databases were collected that would meet variable configuration as continuous speech, several appearances of the person (beard, glasses, hairs), other than front face orientation, moving during recording, two or more speakers in one scene. The reason is that audio-visual databases have problems with time consuming acquisition, storage of the data and their distribution. To acquire a database with high resolution, sufficient number of frames per second and high quality synchronization with acoustic data, needs good quality equipment and storage for the data. With respect to the high demand of the databases they are mainly recorded under different project on individual universities and research labs. Despite of that existing databases contain small number of speakers that do not allowed to develop method which would work in realistic scenario.

With our database we are trying to come closer to realistic scenario. This is done by covering the problem of variable illumination during the speech. The aim of recording database with impaired conditions is to test existing visual parameterization (Císař et al., 2007). This parameterization consist of lip shape information and pixel based information about inner part of the mouth. To stay concentrated on the parameterization it is crucial to preprocess the database. The preprocessing means that database contains information about position of the region of interest (ROI). In this case it is a part of the face around lips. ROI information is used as an input of the visual parameterization algorithm.

This paper is organized as follows. Next section list some of existing databases that were already released. Section three describe database specification and recording. Section four deals with database preprocessing. The last section summarizes the facts about our database.

2. Existing Audio-Visual Databases

Databases can be divided acording to several views. Following subdivision is by type of utterances. First type of database is for recognition of vowels and consonants. We can mention French database (Teissier, 1999), or english one David (Chibelushi et al., 1996).

Another type are databases for recognition of isolated or connected digits. Database Tulip1 (Movellan, 1995) contain 12 speakers that are saying numbers one to four, French database M2VTS (Pigeon and Vandendorpe, 1997) contain all digits for 37 speakers, English one David (Chibelushi et al., 1996) consist of isolated words, newest one is CUAVE from University of Clemon that contains 36 speakers saying numbers(Patterson et al., 2002). The CUAVE and DAVID are the only databases that contain more than one speakers during acquisition.

The last type are databases for continuous speech recognition. They are now enlarging because equipment is being cheaper and the storage of the data is not so problematic as

Database name	Type of database	S	Design criteria	VR	C/M	Utterance	Language
DAVID	alphabet, vowels, words	124	different background	640x480	1/1	47074	English
TULIPS1	isolated digits 1-4	12	segment around mouth	100x75	1/1	50	English
M2VTS	isolated digits 0-9	37	head rotation, glasses, hat	286x350	1/1	unknown	French
CUAVE	isolated, connected digits 0-9	30	hair, glasses, hats	720x480	1/1	7000	English
LPFAV2	continuous mathematical expression	1	no	720x576	1/1	652	Portugese
MANDARIN CHINESE	continuous speech	225	glasses	720x576 768x576	7/12	17000	Chinese
AVOZES	isolated digits 0-9, continuous speech	20	glasses,beard,head rotation	720x480	2/3	unknown	English
X2VTSDB	continuous speech	295	glasses, hair style, head rotation	720x576	2/1	7080	English
IBM-ViaVoice	continuous speech	290	glasses, hair style	704x480	1/1	24 325	English
UWB-05-HSAVC	continuous speech	100	glasses, hair style	720x576	1/2	20000	Czech
UWB-07-ICAVR	continuous speech	50	glasses, hair style, variable light	720x576 640x480	2/2	10000	Czech

Table 1: Comparison of the existing databases. Abbreviation: S represent number of speakers,VR stands for video resolution, C/M number of cameras and microphones, Utterance represent total number of utterances in database.

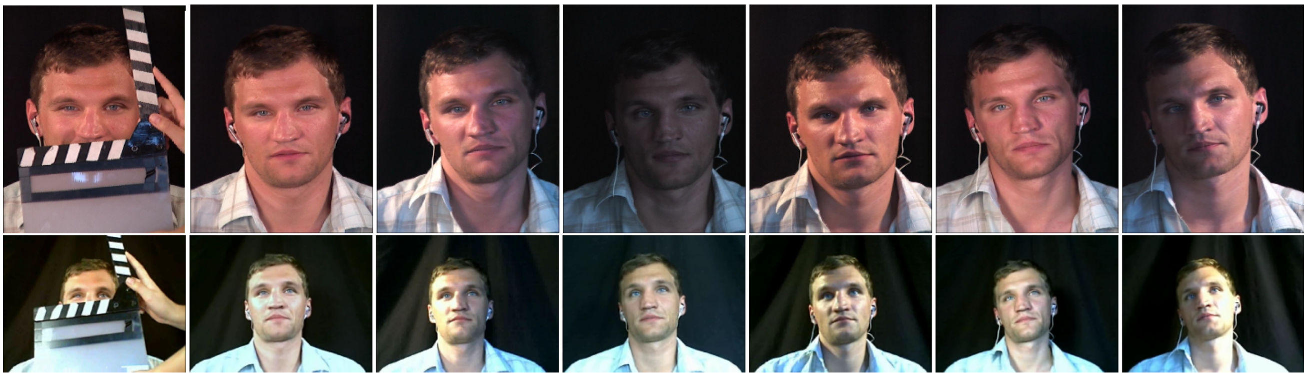

Figure 1: Database recording. First row recorded by VCR camcorder Canon MVX3, the second by webcamera Philips SPC 900NC. First snap is with clapper board, the rest snaps are for illumination conditions IC 1-6 look at the table 2

was in the past. The distribution is also easy thanks to the internet connection. The list of some databases is as follows: LPFAV2(Pera et al., 2004),MANDARIN CHINESE (Liangi et al., 2004), AVOZES (Goecke and Millar, 2004), XM2VTS (Messer et al., 1999), IBM ViaVoice (Neti et al., 2000), UWB-05-HSCAVC (Císař et al., 2005).

There could be many criteria to compare the databases see Table 1 for other comparison.

3. Data Collection

The design criteria for the database is to record data set that would contain variable illumination conditions for a large number of speakers. The visual information of the speech contains part of the vocal tract that can be seen (lips, teeth, tongue and face). It does not contain as much information as acoustic part and it can not be used separately. Rather, it is used as a supplement for acoustic speech recognition. For the visual component of the speech, various parameterizations are used. The overview of existing methods can be found in (Potamianos et al., 2004).

3.1. Database Specification

Database UWB-07-ICAVR (UWB stands for University of West Bohemia, 07 for year of recording, ICAVR for impaired conditions audio visual speech recognition) enlarges the previous database UWB-05-HSCAVC (Císař et al., 2005) which was recorded for research on visual speech parameterization. The purpose to record a new database is to test the existing visual parameterization (Císař et al., 2007) under variable illumination.

Impaired conditions of a visual component of the corpus can be reached by two ways. One way is to artificially process video after recording by various degradation for instance blurring, noise adding or down sampling the resolution. The second one is to change the illumination during recording. This approach is making database more realistic. The variable illumination results in shadows on the face which can not be reached by video post-processing. Other conditions stay optimal during the recording (the head pose is static and capturing is done from the front view).

Impaired conditions of an acoustic component were simulated by playing noise into a headphones to influence the quality of the speech. The acoustic component can be furthermore impaired by adding the noise to acoustic signal. The corpus consists of 50 speakers, 25 males and 25 females. The average age of speaker is 22 years. Each speaker was recorded during one session. Speaker read the text from screen in front of him. After each sentence two seconds long pause was made. Every slip of the tongue was recaptured to obtain fluent speech. The length of an average recording was half an hour. Lights switching was performed 12 times. There were six types of illumination condition. The types can be found in Table 2. The lights switching was repeated twice during one section.

The text of the corpus for each speaker contains 200 sentences. It is divided into two parts. First part has 50 sentences, they are the same for all speakers. The selection was made to cover all phonemes (Radová and Vopálka, 1999). After each eight sentences the illumination was changed. The second part has 150 sentences which are different for each speaker. They are balanced to contain large number of triphones. After each 25 sentences the illumination was changed.

3.2. Database Recording

Database contains visual and acoustic part. Visual part has two recordings. One from VCR camcorder. The second is taken by web camera. Acoustic part was recorded by two microphones, directional microphone and table stand microphone. For synchronization of the data streams the clapperboard was used at the beginning of each recording. During the acquisition the person weared head phones, which produced noise, to influence the quality of a speech. Different illumination conditions were set according to Table 2. The position of the cameras, microphone and lights can be seen from Figure 2.

3.2.1. Acoustic Component of the Corpus

Acoustic part was recorded separately by two microphones connected through amplifier into sound card of desktop PC. First microphone was directional microphone AKG CK55L clip on the front flap, the second one was table stand microphone Sennheiser K6 placed in front of the speaker on left side. For recording program Cool Edit was used and individual setting for each speaker was prepared to obtain the best quality acoustic signal which do not cross the saturation. The format of the data is WAV file with sampling frequency 44KHz, resolution 16bits. The acoustic data for one speaker occupy approximately 200 MB of disk space.

3.2.2. Visual Component of the Corpus

Visual part was recorded by two cameras. The first one was VCR camcorder Canon MVX3i with setting: lens shutter to 1/250 and automatic sharping. The data was stored on the 60 minute tapes and acquired off line through the Firewire interface. The data were encoded in MPEG-4 and saved in AVI format. The example snaps can be seen on top line of the Figure 1. The camera was turned 90 degrees to obtain better resolution of the face. Reason for that is the fact that camera is rather landscape orientated in comparison with face portrait orientation. This reduced the lost of almost 50% of the area of the frame. The camcorder use interlacing which is combination of the two consecutive pictures into one frame. We use AviSynth software to deinterlace the video, by using it we can obtain from the original resolution 720x576 pixels at 25 frames per second the same resolution with 50 fps. During the acquisition were speakers requested not to move their head. Thus the position of the head is static in the whole corpus. The visual data for one speaker from camcorder occupy approximately 5 GB of disk space. The second camera was webcamera Philips SPC 900NC with automatic light balancing settings and resolution 640x480 with 30fps. For on-line acquisition the program VirtualDub was used. The data from web camera was compressed by DivX codec. The camera orientation is landscape. The example snaps can be seen on bottom line of Figure 1. The visual data for one speaker from webcamera occupy approximately 100 MB of disk space. To create the impaired condition the variable illumination during the acquisition was used. Lights used for variable illumination

IC	H1	F1	H2	F2	part1	part2
1	On	On	On	On	8	25
2	On	40%	Off	Off	8	25
3	Off	Off	Off	Off	8	25
4	On	Off	On	Off	8	25
5	Off	On	Off	On	8	25
6	Off	Off	On	Off	10	25

Table 2: Illumination conditions (IC) for database: H1,H2 stands for halogen lights B-7PJ Brilux; L1,L2 stands for digital lights-1000 Fomei; part 1,2 stand for parts of the corpus and it shows the number of sentences taken under each illumination condition. The position of the light can be seen from fig 2.

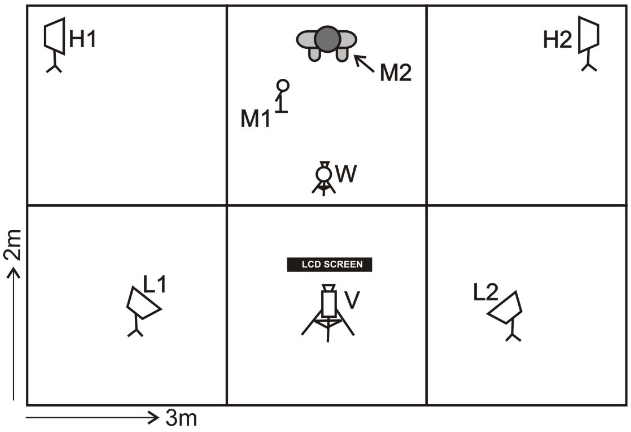

Figure 2: Recording setup. H1,H2 stands for halogen lights B-7PJ Brilux; L1,L2 stands for digital lights-1000 Fomei, W stands for webcamera Philips SPC 900NC, V stands for VCR camcorder Canon MVX3i, M1 stands for table stand microphone, M2 directional microphone placed on the front flap.

was two halogen lights B-7PJ Brilux with 500 Watt light power and two digital lights-1000 Fomei with adjustable power setting 400W, 600W and 1000W. The background of the scene is uniform and black.

4. Database Preprocessing

The whole audio-visual database is annotated and preprocessed. The structure of database for each speaker can be seen from Figure 3.

4.1. Acoustic Annotation

Start point (clapper board) of the session, with each speaker, was manually labeled at high accuracy (milliseconds). All mistakes (e.g. split of the tongue, word missspeling) during acquisition were recaptured. Each sentence is now being annotated. The actual pronunciation is transcribed into the phonetic notation included notes about features like breath and other noises that appeared during the acquisition.

4.2. Visual Annotation

Synchronization of the acoustic and visual part is done by finding a clapper board at the beginning of the video file. The time markers from annotation of acoustic signals are used to split the video recordings in separate sentences. All frames are accompanied with visual label. The database should be used for visual parameterization testing for this reason the position and size of the region of interested (ROI) is crucial. ROI is the part of the face where are the lips and area around them, see Figure 3. The processing of the video files is semi-automatic. Several frames from the video is chosen to train the algorithm active shape model (ASM) (Cootes and Graham, 1995). We first assume to use adaboost, but due to a variable lighting the results were not promising. In cases, where face is under illumination conditions IC2 or IC6 (see Table 1), a mouth is not found properly. Instead ear is found. ASM uses differential profile that is why it is not influence by illumination changes. The size of ROI is set according to distance between eyes. Folder description include text file which contain 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 preprocessing of the database. This procees is controlled by annotater. He can see the results of algorithm, when the position of the mouth differ from previous frame algorithm stop and ask for verification of the ROI position.

5. Conclusion

This paper present design and recording of the new czech audio-visual speech database intended for experiments on various impaired conditions. We concentrated on changing the illumination. Result is audio-visual speech database UWB-07-ICAVR which is a valuable resource for testing algorithm for visual speech parameterization under variable illumination.

6. Acknowledgements

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416 and by the Ministry of Education of the Czech Republic, project No. ME08106.

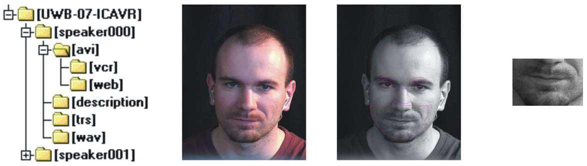

Figure 3: right side: Database structure. Avi folder contains both visual recording (camcorder and webcamera), description contain information about region of interest (ROI), trs contain transcription of acoustic part, wav contain acoustic files left side: original frame, eyes and mouth find by ASM, selected ROI

7. REFERENCES

C.C Chibelushi, S. Gandon, and J. S. D. Mason. 1996. Design issue for a digital audio-visual integrated database. In *Integrated Audio-visual Processing for Recognition, Synthesis and Communication*.

P. Císař, M. Železný, Z. Krňoul, J. Kanis, J. Zelinka, and L. Müller. 2005. Design and recording of czech speech corpus for audio-visual continuous speech recognition. In *Auditory-Visual Speech Processing Workshop 2005*.

P. Císař, M. Železný, J. Zelinka, and J. Trojanová. 2007. Development and testing of new combined visual speech parametrization. In *International Conference on Auditory-Visual Speech Processing*.

C.J.and Cooper D.H Cootes, T. F.and Taylor and J. Graham. 1995. Active shape models - their training and application. In *Computer Vision and Image Understanding*, volume 61, pages 38–59, January.

R. Goecke and J.B. Millar. 2004. The Audio-Video Australian English Speech Data Corpus AVOZES. In *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004*, volume III, pages 2525–2528, Jeju, Korea, October.

Luhong Liangi, Yu Luo, Feiyue Huang, and A.V. Nefian. 2004. A multi-stream audio-video large-vocabulary mandarin chinese speech database. In *IEEE International Conference on Multimedia and Expo*, volume 3, pages 1787 – 1790.

K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre. 1999. Xm2vtsdb: The extended m2vts database. In *Audio- and Video-based Biometric Person Authentication, AVBPA'99*, pages 72–77. Washington, D.C., March 1999. 16 IDIAP–RR 99-02.

J.R. Movellan. 1995. Visual speech recognition with stochastic networks. In G. Tesauro, D. Toruetzky, and Leen T., editors, *Advances in Neutral Information Processing Systems*, volume 7. Cambridge, MIT PRESS.

C. Neti, G. Pontamianos, J. Luettin, and I. Matthews. 2000. Audio visual speech recognition. Technical report, Center for Language and Speech Processing.

E. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy. 2002. Cuave: a new audio-visual database for multimodal human-computerinterface research. In *Acoustics,* *Speech, and Signal Processing*, volume 2, pages 2017 – 2020.

Vitor Pera, Antonio Moura, and Diamantino Freitas. 2004. Lpfav2: a new multi-modal database for developing speech recognition systems for an assistive technology application. In *SPECOM*.

Stéphane Pigeon and Luc Vandendorpe. 1997. The m2vts multimodal face database (release 1.00). In *AVBPA '97: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 403–409, London, UK. Springer-Verlag.

G. Potamianos, C. Neti, J. Luettin, and I. Matthews, 2004. *Audio-Visual Automatic Speech Recognition: An Overview*, chapter 10. Issue in Visual and Audio-Visual Speech Processing.

V. Radová and P. Vopálka. 1999. Methods of sentences selection for read-speech corpus design. In *TSD '99: Proceedings of the Second International Workshop on Text, Speech and Dialogue*, pages 165–170, London, UK. Springer-Verlag.

J.and Schwartz J.-L.and Guerin-Dugue A. Teissier, P.and Robert-Ribes. 1999. Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing*, 7:629 – 642.