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 in Pilsen, Czech Republic
{trojana, mhruz, campr,zelezny} @kky.zcu.cz

INTRODUCTION

Audio-Visual database UWB-07-ICAVR is presented here.

This database was recorded for evaluation of existing visual parameterization.Visual information is used for better performance of speech recognition under noisy environment.

It enlarges the database UWB-05-HSCAVC which was recorded for development of visual parameterization

The difference between the databases is that visual part of the new database is impaired with different illumination conditions see database specification for details

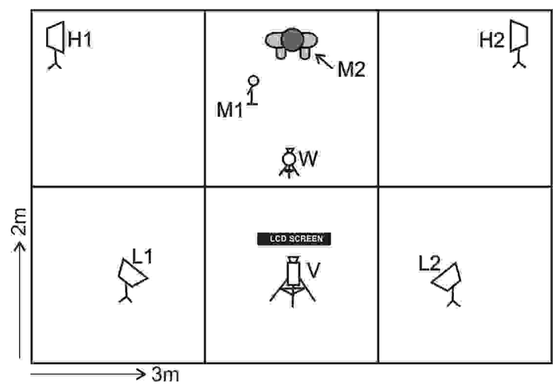

Figure 1: Recording setup. H1, H2 stands for halogen lights B-7PJ Brilux; L1, L2 stands for digital lights 1000 Fomei; W stands for web camera Phillips SPC 900NC; V stands for VCR cam-corder Canon MVX3i; M1 stands for table microphone; M2 is directional microphone placed on the speaker chest.

DATABASE SPECIFICATION

Impaired conditions for visual part of the corpus can be obtained by two ways

- artificially process videos (downsampling, bluring, adding noise)
- simulated light conditions (different source of light that change intensity of picture and produce shadows on face).

We have used six types of illumination during the acquisition, the position of the lights can bee seen on Fig 1, and specification of variable illumination can be seen in Tab 1.

Since the visual parameterization that is going to be tested works on front view only, we kept the head pose during the acquisition in static frontal pose

We have also impaired acoustic component:

- during the acquisition we played noised into the headphones to influence the speech of the speaker
- added noise to recorded acoustic part

Totally 50 speakers were recorded. Half of them were men. All speakers were reading 200 sentences (first 50 were same - it will be used as testing data, rest 150 were different - it will be used as training part)

IC	H1	L1	H2	L2	test part	tran. part
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	8	25

Table 1: Illumination conditions (IC) for database: H1, H2 stands for halogen lights; L1, L2 stands for digital lights. Test part and training part shows numbers of sentences under each illumination condition.

DATABASE RECORDING

Visual part was recorded by two cameras: VCR camcorder Canon MVX3, web camera Phillips SPC900NC

Acoustic part has also two parts, One stream was recorded by directional microphone, the second was table stand microphone.

Position and orientation of the equipment can be seen from Fig1.

The recorded data of one section for one speaker contains:

- 5 GB for visual part
- 0.2GB for acoustic part

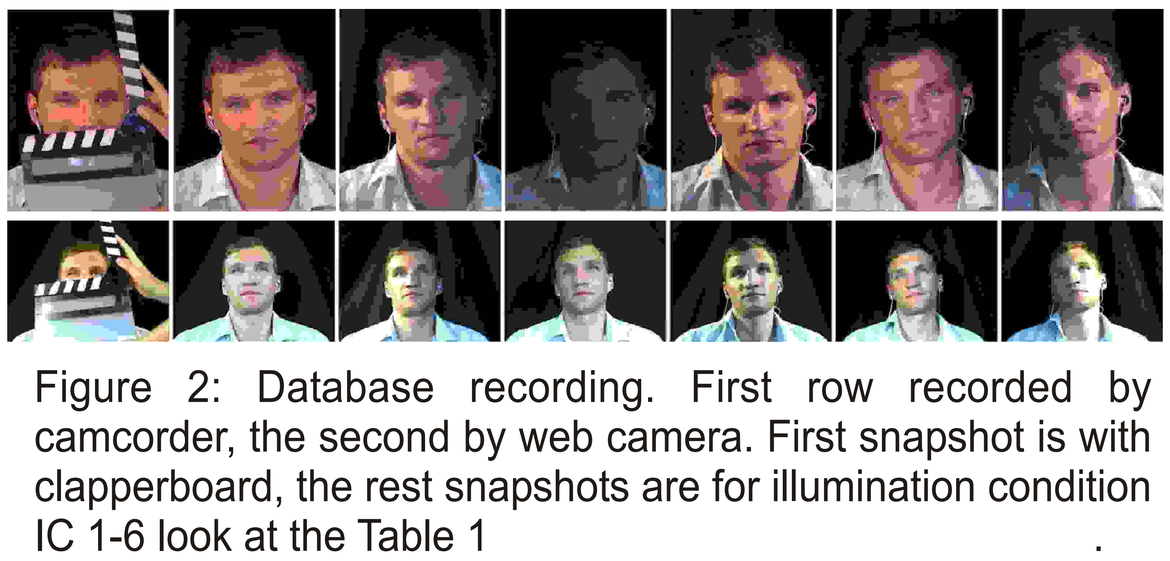

Figure 2: Database recording. First row recorded by camcorder, the second by web camera. First snapshot is with clapperboard, the rest snapshots are for illumination condition IC 1-6 look at the Table 1 .

PREPROCESSING & CONCLUSION

The database contains four separate streams (two visual and two acoustic). Synchronization is done by a clapper-board.

The purpose of the database is to test visual parameterization for speech recognition. Region of the mouth has to be specified in the description that goes with database.

The description file contains coordinates of a mouth, its rotation and size of the region of interest.

Annotation of the speech is made in Transcriber. Time markers from annotation are used for splitting the videos into separate sentences

Audio-visual database UWB-07-ICAVR is a valuable resource for testing algorithms for visual speech parameterization.

Figure 3: a) Database structure. AVI folder contains both visual recording (camcorder and web camera), description contains information about region of interest (ROI), TRS contains transcription of acoustic part, wav contains acoustic files; b) left site: eyes and mouth found by adaboost; right site: selected ROI

Acknowledgment: 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.

lrec 2008 marrakech