Multi-Modal Emotional Database: AvID

Rok Gajšek, Vitomir Štruc and France Mihelič
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, SI-1000 Ljubljana, Slovenia
E-mail: {rok.gajsek, vitomir.struc, france.mihelic}@fe.uni-lj.si, http://luks.fe.uni-lj.si/en/index.html

Anja Podlesek, Luka Komidar, Gregor Sočan and Boštjan Bajec
Psychological Methodology
Faculty of Arts
University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
E-mail: {anja.podlesek, luka.komidar, gregor.socan, bostjan.bajec}@ff.uni-lj.si

Keywords: multimodal database, speech recognition, emotion recognition

Received: November 5, 2008

This paper presents our work on recording a multi-modal database containing emotional audio and video recordings. In designing the recording strategies a special attention was payed to gather data involving spontaneous emotions and therefore obtain a more realistic training and testing conditions for experiments. With specially planned scenarios including playing computer games and conducting an adaptive intelligence test different levels of arousal were induced. This will enable us to both detect different emotional states as well as experiment in speaker identification/verification of people involved in communications. So far the multi-modal database has been recorded and basic evaluation of the data was processed.

Povzetek: V članku je predstavljeno snemanje večmodalne zbirke, ki vsebuje audio in video posnetke različnih čustvenih stanj.

1 Introduction

This study paper describes initial attempts to collect a multimodal emotional speech database as a part of our research under the ongoing interdisciplinary project *"AvID: Audiovisual speaker identification and emotion detection for secure communications"*. The goal of the project is to use speech and image technologies in video telecommunication systems for identification/verification and detection of the emotional state of persons involved in communication. Such a system should provide additional information about the identity and psychophysical condition displayed on the communication devices enabling more secure and credible exchange of information. Project partners are from the Faculty of Electrical Engineering – Department for Automatics from the University of Ljubljana and Jožef Stefan Institute – Department for Intelligent Systems from Ljubljana, the Faculty of Arts – Department for Psychology from the University of Ljubljana and the industrial R&D company Alpineon D.O.O. from Ljubljana.

Unfortunately most of the available speech databases with emotional speech were obtained by recording of acted different type of emotions usually from professional actors [5, 6, 11, 2, 3, 7][1] and therefore do not represent very adequately data for training and testing procedures for speaker psychophysical condition detection in real environment. To distinguish between normal and non-normal condition, and to compare speaker verification performances we also need quite a lot of speech material with normal speech, that is also usually not the case for available databases. Although we plan to perform our experiments on as much available data as possible, we also intend to obtain some data providing more realistic conditions for our task. Therefore we are planing to collect reasonable amount of audio and video recordings of spontaneous speech in normal (relaxed) and non-normal psychophysical conditions (the conditions of excitement and arousal with different valence, both positive and negative) from a representative group of speakers. Initial strategies to obtain desired speech corpora and recording setup along with the statistics of already recorded data are described in the following sections.

2 Recording strategies

In the beginning, each participant was told that the main purpose of the experiments was to examine whether dif-

[1] There are of course some exceptions as, for example, the German SmartKom database [12] which was recorded using a Wizard of Oz technique and tried to evoke different emotions in the participants.

ferent measures of his/her state could be used in an adaptive test of intelligence. The biometric measures to be indicative of his/her psychophysical state at different moments were: psychophysiological response (the electrodermal and electrocardiographic response), verbal response, and facial expression. After a written consent to participate in the study was obtained from each individual, sensors were placed on the index and the middle finger on the left hand and the audio and visual recording started. The participant was instructed to speak loudly enough and not to move. With the right hand he/she had to hold a computer mouse in order to prevent the hand from excessive movement.

To obtain recordings of speech in both neutral and changed psychophysiological state of the participant, we designed an experiment composed of four parts. In Part I, after the participant introduced himself/herself with a few words (stated the name, the place of living, age, and main occupations), photographs with neutral content were presented on the screen. The participant was instructed to describe each photograph in detail, as if he/she were describing what he/she sees to a blind person. In this part we supposedly measured his/her verbal fluency. When he/she finished with the descriptions, he/she instructed the experimenter to continue with the presentation of the next photograph.

Before the start of Part II, the participant was told that we will be assessing the efficiency of his/her verbal instructions given to a teammate in order to achieve a common and specific goal. We explained to the participant that the team, which will involve himself/herself and the experimenter, will play a computer game (Tetris) and that he/she will observe the progression of the game on the computer monitor and will be giving verbal instructions, whereas the experimenter will not be able to observe the game and will carry out his/her orders by pressing the appropriate buttons on the keyboard. If the participant had no prior experience with the game, we explained the rules of Tetris and let him/her play for a few minutes. The participant who observed the ongoing game on the screen had to lead the experimenter through the game by uttering the following four commands: Left ('Levo' in Slovene), Right ('Desno'), Around ('Okrog'), and Down ('Dol'). The goal of the team was to achieve the highest score possible. Passive commands (e.g. 'Around' instead of 'Turn it around') were chosen in order to be suitable for use also in Part III of the experiment. At the end of the game the participant had to tell the experimenter what score they had achieved, what happened during the game, and why the game ended.

The ongoing game was recorded by CamStudio screen capture program. In Part III the recorded movie of the game was played and the participant had to describe what was happening on the screen by using the same four commands as in Part II. At the end the same description of the events on the screen had to be given as at the end of Part II. The aim of Part II was to obtain positive arousal (joy, satisfaction) as well as negative arousal (frustration, anger), whereas Part III was carried out to obtain exactly the same utterances in a relaxed, non-aroused state, because in Part III the participant was just a passive observer.

At the beginning of Part IV, the participant was told that he/she will be given an adaptive intelligence test where the difficulty of the task will be chosen by the computer according to (i) the correctness of the answer in the previous task, (ii) the mental strategy used for solving the task, and (iii) the biometric measures (EDR and heart rate). We explained to her that several values will be presented on the left part of the screen: the momentary IQ value, the arrows pointing upwards when the IQ estimate was increasing and downwards when it was decreasing, the momentary values of EDR and EKG measures, and the time that remained for solving the current task. On the right part of the screen, matrices with different figures or symbols were presented with one element absent. The participant had to reason aloud about the principles of the arrangement of matrix elements in rows and columns and find the proper solution among five to six possible answers. The participants believed they had to reason aloud so that the experimenter will be able to assess their mental strategy used for solving the task. After the experimenter showed two examples of matrices and explained how the reasoning should be verbalised, 20 matrices had to be solved, some of which were very difficult or did not have a known solution. If the participant ceased to speak aloud, the experimenter encouraged her to verbalise her thoughts. After the solution was found, the experimenter clicked some buttons to input the chosen solution and the presumed category of mental strategy. He could choose among six options with which he controlled the changes in the unfounded IQ estimate. He raised the IQ value only when the correctness of the solution was obvious, the reasoning was straightforward and solution was derived quickly. In other cases the IQ value was decreased. The main purpose of decreasing the IQ value was to increase the participant's subjective stress level. Besides decreasing the temporary IQ value, the experimenter could also manipulate participant's stress level by increasing the EDR and heart rate values. In order to attract attention to the EDR and heart rate indicators during the test, the values changed its colour from black to red when a certain value was exceeded.

After Part IV was over we debriefed the participant. The experimenter explained that the temporary and final IQ scores were not valid estimates of her intelligence and that the real aim of the study was to obtain the recordings of speech in the normal, relaxed state and in the aroused, stress-induced emotional state. The participants were then asked to describe (freely) their feelings, thoughts, and involvement in each part of the experiment. In the end, some general data on participants and their speech characteristics were gathered (see Table 1).

Subject	Sex	Age	Voice type	Health	Smoking	Overall mood	Dialect	Speech peculiarities
01	M	20	baritone	normal	NO	slightly tense	Central	
02	F	37	mezzo-sop.	cold	casual	relaxed	Eastern	
03	F	19	mezzo-sop.	normal	NO	relaxed	Littoral	
04	F	25	mezzo-sop.	normal	NO	relaxed	Eastern	
05	F	21	mezzo-sop.	normal	NO	relaxed	Central	
06	F	26	mezzo-sop.	normal	YES	NA	Central	
07	M	21	bass	normal	NO	relaxed	Central	rash speech
08	F	19	soprano	normal	stopped	distracted	Eastern	
09	F	20	mezzo-sop.	normal	NO	distracted	Littoral	
10	M	28	tenor	normal	NO	relaxed	Central	
11	F	26	mezzo-sop.	normal	NO	relaxed	Eastern	
12	F	20	mezzo-sop.	normal	NO	relaxed	Lower Car.	
13	F	27	mezzo-sop.	normal	NO	relaxed	Eastern	
14	F	20	mezzo-sop.	normal	YES	relaxed	Eastern	
15	F	27	soprano	normal	NO	relaxed	Central	

Table 1: Basic participants' data relevant for recorded speech analysis.

3 Recording conditions and inventory

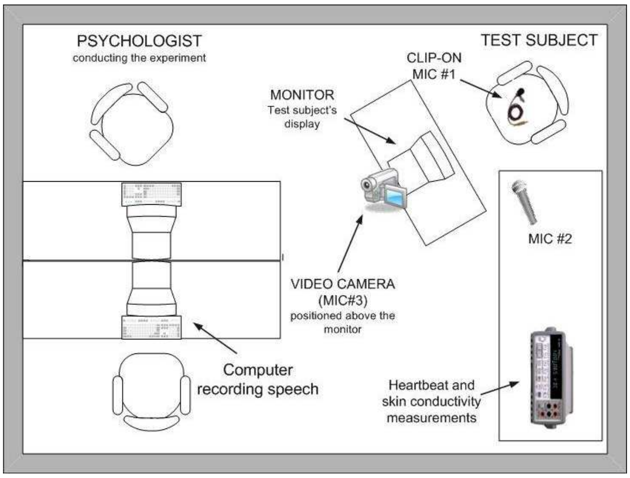
PSYCHOLOGIST
conducting the experiment
TEST SUBJECT
CLIP-ON MIC #1
MONITOR
Test subject's display
MIC #2
VIDEO CAMERA (MIC#3)
positioned above the monitor
Computer recording speech
Heartbeat and skin conductivity measurements

Figure 1: An outline of the recording setup.

Recordings were done in a closed room using a digital video camera and three microphones [10]. We used several microphones as in our previous GOPOLIS database [10] to enable some environment and channel normalisation tests. For the reference also some physiological sensors were used to detect changes in the heart beet rate and skin conductivity during the tests on few test objects.

Each participant was asked to position himself/herself in front of the computer monitor shown in the upper right corner of Fig. 1. Behind the monitor a digital camera mounted on a tripod was placed to capture the video recording (i.e., video as well as one channel of audio data). To ensure an appropriate quality of the captured video data the participant was seated in front of a relatively homogeneous, white background and a light source was directed towards the participant's face. This setup resulted in the recorded video sequences showing a fairly "clean", i.e., without to much shadows, frontal view of the participant's face - see Fig. 2 where the recording setup is presented.

Note that the quality of the recorded video data could have been further improved by using additional light sources or diffused light, however, as our goal was to collect a realistic database the employed setup fully sufficed for our requirements.

For capturing the audio signal two microphones were used in addition to the one integrated in the digital camera. The first, denoted as MIC #1 in Fig. 1, was attached on the participant's clothing near the chest, while the second, denoted as MIC #2 in Fig. 1, was positioned on the nearby desk. Both microphones were hooked to a computer which was used for recording and storing of the audio data.

Two people were supervising the acquisition of the database: (i) a psychologist who was in charge of the recording session and tried to induce a "non-normal" psychophysical condition in the participant by applying the strategies presented in Section 2 and (ii) a technician who overlooked the technical aspects of the acquisition process.

3.1 Video data collection

The video part of the AvID emotion database was acquired using a high-definition Sony HDR-SR11E digital handycam which captures video at a resolution of 1920 × 1080 pixels and a bit-rate of 16Mb/s. The video data was recorded at a frame aspect ratio of 16 : 9 and later on archived in the AVCHD (Advanced Video Codec High Definition) format[2]. Specifications of the employed format

[2] A high-definition format jointly established by Panasonic and the Sony Corporation.

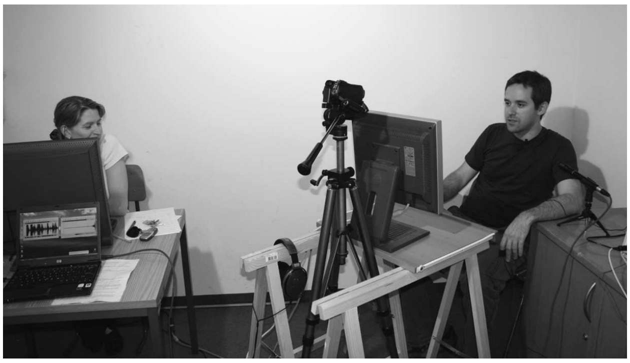

Figure 2: The recording setup.

for the video data can be found in Table 2.

Video signal	$1080/50i$
Pixels	1920×1080
Aspect ratio	16 : 9
Compression	MPEG4 AVC/H.264
Luminance sampling frequency	74.25 MHz
Chroma sampling format	4:2:0
Quantization	8-bit

Table 2: Specifications of the employed AVCHD format.

A high-definition camera was chosen for capturing the video data of the database for several reasons: (i) different kinds of experiments (e.g., biometric verification or identification, emotional state recognition, lip reading, etc.) can be performed on high quality video, (ii) with simple image- and video-processing techniques the quality of the video data can easily be degraded and research can be conducted on lower-quality video, and (iii) as high-definition technology is spreading with an increasing speed, it will soon find its way into peoples daily lives; with its widespread deployment the technology will also become easily affordable and, therefore, suitable for employment in low- (or medium-) cost recognition systems. A sample frame captured with Sony's HD camera in the AVCHD format is shown in Fig. 3.

3.2 Audio data collection

As mentioned above the audio signal was captured using three different microphones. Channel number one was recorded using a Sennheiser ew122-p G2 system with a clip-on microphone which transmitted the signal to the recording computer via radio waves. The microphone was pinned to the speakers chest roughly $10 - 20$cm away from the speaker's mouth.

The second channel was captured with a Shure PG81 microphone which was positioned approximately $30 - 40$ cm away from the speaker. Both microphones specify a frequency range of $40 - 18000$ kHz. Channels were recorded at a sampling rate of 16 kHz and 16-bit linear encoding.

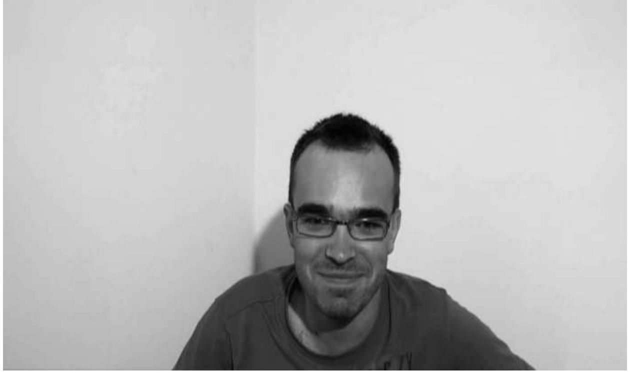

Figure 3: A sample frame from the AvID audio–video emotional database.

The third channel was acquired from the employed video camera's built in microphones that record in Dolby Digital 5.1 and use AC-3 compression for audio storage.

4 Database description

A total of 15 native Slovenian speakers (12 female and 3 male) were recorded with one session lasting approximately an hour. After extracting only the speaker's speech from the session we got roughly a half an hour of usable audio per speaker. For the video part of the database one continuous recording was captured for each of the participants resulting in over 15 hours of high-definition video. As already mentioned in the previous section the participants were recorded in front of a white background and with frontal illumination. The average inter-ocular distance, which is the traditional measure of the size of the face in an image or video frame, is more than 150 pixels. Similar databases (uni- or multi-modal) used either for assessing biometric identification/verification or emotion recognition algorithms, such as the XM2VTS [9], the Cohn-Kanade [4] and the eNTERFACE'05 [8] databases, typically feature face images (or video sequences) with a distance of $40 - 60$ pixels between the left and the right eye. The AvID database is therefore suitable as the foundation for the development of recognition algorithms that make use of high resolution information.

The audio recordings were later split to shorter utterances - approximately one utterance per sentence. Transcriptions and labels describing emotional state of the speaker were made using Transcriber tool [1] and followed the LDC broadcast speech transcription conventions[3].

[3] LDC broadcast speech transcription conventions: http://projects.ldc.upenn.edu/Corpus_Cookbook/ transcription/

5 Evaluation results

Subjective reports were analysed - descriptions of the states at the beginning of the experiment and within each part of the experiment were classified into five categories (see Figure 1) where possible. Where the category is composed of two arousal levels (e.g., moderately and highly aroused), the first one reflects the prevalent state and the second reflects temporary peaks of slightly elevated stress level. Mostly the participants reported of the relaxed or slightly tense state prior to the experiment (when the sensors were attached and the procedure was explained to them). In Part I, when describing photographs, most of them reported no tension, and some reported a slight tension that later vanished. Their arousal was increased slightly while playing Tetris. Some participants reported negative emotions (e.g., irritation) in Part III due to the inability to take over the control. The majority reported that Part IV was difficult and stressful because they had troubles with verbalising their reasoning and were worried and puzzled about the calculated IQ value. This was reflected in a notable decrease of speech loudness. The change towards higher arousal during the experiment is indicated with the prevalence of darker pattern in Figure 4, whereas the transparent patterns represent a more relaxed state. It may be concluded that the experimental situations elicited the presumed levels of arousal: neutral emotional state with describing photographs and events, and arousal in playing an exciting computer game and in the situation where an individual wants to perform well under social pressure.

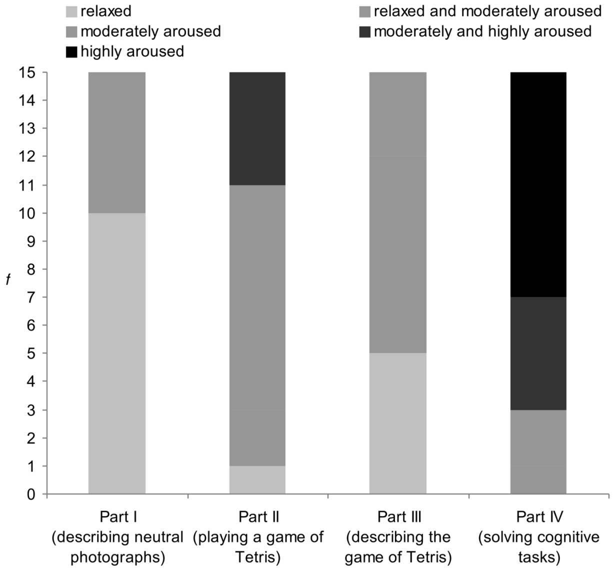

Figure 4: Levels of arousal during different parts of the experiment based on the participants' subjective report.

5.1 Future prospects

In future studies, a triangulation of different methods will be used to assess the arousal and the emotional state of the participants more systematically and objectively. To obtain additional indicators of the participant's emotional state we will use: (i) standardised instruments of subjective emotional experience, (ii) the psychophysiological measures, such as EDR and EKG, and (iii) behavioural expressions.

6 Conclusion

The goal of recording a multi-modal speech database containing different spontaneous emotions was achieved. Due to well selected experiments different levels of arousal were induced and measured by different biometric parameters: facial expression - video, verbal response - audio and psychophysical response - electrodermal and electrocardiographic response. Video and audio comprise the database, where psychophysical measures are only used to provide an objective information about the level of arousal.

Enough data was collected (which is especially important for speech research) to form a bases for future studies on speaker identification/verification, emotion recognition and spontaneous speech analysis research.

Acknowledgement

This work was supported by the Slovenian Research Agency (ARRS), development project M2-0210 (C) entitled "AvID: Audiovisual speaker identification and emotion detection for secure communications."

References

[1] C. Barras, E. Geoffrois, Z. Wu and M. Liberman (2001) Transcriber: development and use of a tool for assisting speech corpora production, *Speech Communication*, pp. 5–22.

[2] A. Battocchi and F. Pianesi (2004) DAFEX: Un Database Di Espressioni Facciali Dinamiche, *Proceedings of the SLI-GSCP Workshop "Comunicazione Parlata e Manifestazione delle Emozioni"*, pp. 1–11.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss (2005) A Database of German Emotional Speech, *Proceedings Interspeech 2005*, pp. 1–4.

[4] T. Kanade, J.F. Cohn and Y. Tian (2000) Comprehansive database for facial expression analysis, *Proceedings of the 4th AFGR'00*, pp. 46-53.

[5] LDC (1999) SUSAS (Speech Under Simulated and Actual Stress), *Proceedings of the 4th AFGR'00*, Language Data Consortium,

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78.

[6] LDC (2002) Emotional Prosody Speech and Transcripts, Language Data Consortium, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28.

[7] O. Martin, I. Kotsia, B. Macq and I. Pitas (2006) The eNTERFACEŠ05 Audio-Visual Emotion Database, *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp. 1–8.

[8] O. Martin, I. Kotsia, B. Macq and I. Pitas (2006) The eNTERFACE'05 Audio-Visual Emotion Database, *Proceedings of the 22nd International Conference on Data Engenireeng Workshops*, IEEE Computer Society.

[9] K. Messer , J. Matas, J. Kittler, J. Luettin and G. Maitre (1999) XM2VTSDB: the extended M2VTS database , *Proceedings of AVBPA'99*, pp. 72–77.

[10] F. Mihelič, J. Žganec Gros, S. Dobrišek, J. Žibert and N. Pavešić (2003) Spoken language resources at LUKS of the University of Ljubljana, *Int. J. Speech Technology*, pp. 221–232.

[11] V. Hozjan, Z. Kačič and B. Horvat (2001) Prosody feature analysis for emotion modeling, *Electrotechnical Review*, pp. 213–218.

[12] U. Turk, (2001) The Technical Processing in SmartKom Data Collection: a Case Study, *Proceedings of Eurospeech*, pp. 1541–1544.