A Review on Creation and Structure of Emotional Multimodal Databases

Vedran Strčić, Lucia Načinović, Ivo Ipšić
Department of Informatics
University of Rijeka
Omladinska 14, 51000 Rijeka, Croatia
{vstrcic, lnacinovic, ivoi}@inf.uniri.hr

Abstract. *This paper presents an overview of emotional multimodal databases (EMD) and the resulting corpus data. EMD are the basis of systems that will register and respond to emotion. Construction of multimodal corpus is a complex process that consists of several steps including subject selection, setting up and managing data acquisition scenario and various data postprocessing tasks. The paper overviews some of the existing EMD, adresses the problems commonly associated with their application and indicates the future directions for the development of EMD for Croatian language.*

Keywords. multimodal database, multimodal corpus

1 Introduction

Emotion is one of the key factors in communication [15]. So far, all communication between human and the machine has mostly been limited to series of simple instructions and result feedback. Communication as such does not always convey the message adequately, consequently the system may not always realize its purpose fully and efficiently.

To improve human-machine communication systems should have the ability to handle, besides audio and visual information, user's emotions as well. A multimodal database is necessary for the development of a system which will utilize recognition of spoken language and visual information (gestures, facial expresions) to enhance the quality of interactions [1]. The creation of such database is a complex process, the first step of which is deciding on the database specifications. To specify those parameters we need to clarify what information we want to collect and the way we will collect it [1].

In this paper we present an overview of emotional multimodal databases and the process of their creation. Understanding this process is the first step towards the development of an emotionally intelligent system.

In section 2, the terms of emotion and emotionally intelligent system are defined along with the description of the purpose of a multimodal database. Section 3 describes how emotions are obtained, processed, and stored to a database. The next two sections give an overview of some of the existing EMD and EMD tools and finally some conclusions are presented in section 6.

2 Description and definition of terms

One of the challenges in building emotionally intelligent systems is the automatic recognition of affective states [15]. Humans recognize those states by considering information context and speaker's individual traits while observing his bodily reactions. Obtaining and assessing all that information presents a challenge for a computer system [15].

To be able to develop an emotionally intelligent system that by definition has an ability to perceive, interpret, express and regulate emotions [2], we should first have a basic understanding of what emotion is. Most authors agree that emotion is a multifaceted phenomenon which encompasses a diversity of processes, such as appraisal, facial ex-

pressions, bodily responses, feeling states, action tendencies, or coping strategies [3].

Looking at the concept of an emotion we can now extrapolate how the representation of such phenomenon will look like in our system and anticipate the type of data we will have to work with to create an adequate digital representation.

To categorize an emotion we can use dimensional approach [20]. According to this approach, all emotions are categorized by arousal and valence, and we can assign them to a certain position in a two dimensional plane. In figure 1, we can see an example of assigning emotions associated with various images to a plane [18]. Each dot in this plane represents one emotionally annotated picture.

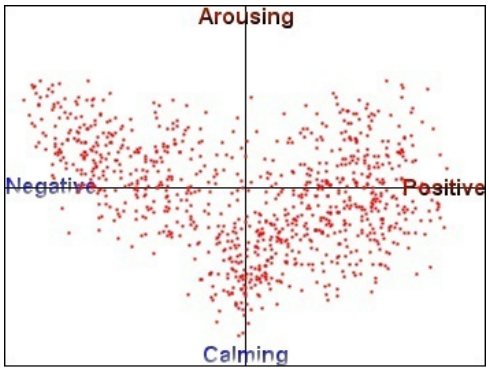

Figure 1: An example of two dimensional plane holding categorized emotional data

To store and organize all multimodal interactions a multimodal database is used. Its objectives are to share the knowledge about specific types of emotional information that are used in interactions and to collect adequate quantity of multimodal data that will be used to train and test the modules of the system. Different data, such as speech, video, heart rate and other, are stored in complimentary channels and synchronized [1]. Organized multimodal corpus is a base for modeling procedures of emotionally intelligent system.

3 Creation of Emotional Multimodal Databases

First step in the process of creation of multimodal database is specifying how to obtain emotional information, or more specifically how to encourage subjects to act emotionally [4]. Looking at this criteria there are three different types of multimodal databases we can get: acted emotions, natural spontaneous emotions and elicited emotions.

3.1 Acted opposed to natural emotions

The most expressive results are normally acquired with acted emotions, while with spontaneous and elicited emotions more natural results are obtained [5]. Problem with acted emotions is inadequate representation of real environments [5], however they are easier to obtain so we can generally collect more training data this way. This process usually involves professional actors performing their roles according to a set plan.

Natural emotions are harder to obtain due to necessity for creation of environment where actors are not aware of their true role. The best results have mostly been obtained by creating scenarios where actors have been aware of being recorded, but not knowing that it was their emotions that have been observed, so they were still expressing them naturally.

Example of this is process of creating the AvID multimodal database [5] where in one of the stages participants were told that efficiency of their verbal instructions to a teammate will be assessed. Participant was instructed to look at computer screen and give verbal instructions to a teammate (experimenter) who was playing a game of Tetris without seeing the screen. The goal was to obtain both positive and negative arousal depending on their success or failure.

Physiology-Driven Adaptive Virtual Reality Stimulation for Prevention and Treatment of Stress Related Disorders [19] is an example of specificaly aimed system that deals with elicited emotions. It's purpose is making soldiers more resilient to adverse psychological effects of combat and healing those who already suffer from combat-related psychological disorders. Specifically, the virtual stimuli of anxiety-provoking situations are delivered gradually to the patient and the system provides an automated assessment of the patient's emotional state based on continual interpretation of multiple acquired physiological signals.

The three major tasks of this system [19] are: time-syncrhronized generation of emotionally and semantically aligned multimedia stimuli, estimation of person's emotional states based on the acquired physiological response, and adaptive closed-loop control that leads to subsequent generation of new stimuli. The stimuli can be provided in various media forms, like static pictures, sounds and video clips.

3.2 Recording environment

Second step, after selecting one of the strategies and choosing the actors, is securing peaceful and isolated recording environment and setting up the necessary equipment. Such environment is necessary so the background noise would be avoided and good lighting ensured [5] [7]. Goal of this step is to make sure the highest possible quality audio and video data is obtained.

Equipment that is normally used is at least one microphone and a video camera if both audio and video data are being obtained, and possibly other measurement devices like heart rate and skin conductivity measurement units and other physiological sensors. PC with software that synchronizes the recording of data from all sources and an adequate permanent memory medium for storage are also needed [4] [5] [6] [7].

After environment is set up audio and video recording settings have to be adjusted. This is an important factor to be considered. Due to equipment having its limitations a compromise has to be made between recording quality and processing power. Possibly large resulting data size also needs to be considered [7].

3.3 Annotating the data

After recording is completed the speech is transcribed and all acquired data is annotated. Annotation is a process of describing the data giving it a meaning. Each modality has to be annotated in order to enable subsequent computation of the relevant parameters of emotional behavior [8].

Annotation is normally done by appropriate software applications some of which are Anvil, ELAN, EXMARaLDA, SLAT, PALinkA, Praat, Wavesurfer and other [9] [10]. For the speech transcription both verbal and non verbal sounds are annotated [8]. Non verbal sounds include pauses, breaths, etc. Besides segments of waveforms and their corresponding words, prosodic events, accents and phrasing can be labeled as well. It is the first step in obtaining speech intonation models. In figure 2, an annotated Croatian sentence with corresponding waveform, pitch contour, labels for beginnings and endings of segments and intonation labels can be seen.

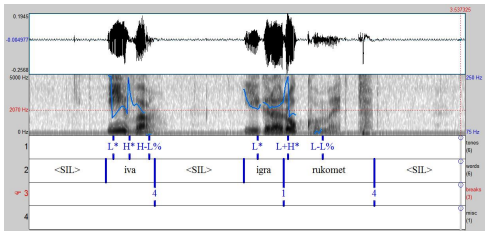

Figure 2: A screen shot of Praat working with speech transcription

In videos the important parts of human body are annotated, those are normally face expressions, head movements and hand gestures [8]. Emotional meaning is finally assigned to each expression and gesture.

4 Overview of EMD tools

As we described in section 3, after recording is completed raw data has to be transcribed and annotated so a meaning is assigned to it. Several tools that serve that purpose have become available [16] [17]. Here is a short overview of some of them.

Praat (v4.0.43) [17] is a phonetics tool used for speech analysis and synthesis. After importing an audio file, user sequentially marks beginnings and endings of segments on the waveform and types in the words for those segments. The result is text file containing transcribed text together with time stamp info. Praat can run on both Windows and Linux and it offers an easy segmentation interface, however it can not handle overlapping speech, when several speakers speak simultaneously, and the visibility of large segments is somewhat impaired.

Transcriber (v1.4) [17] is a transcription tool that runs on both Unix and Windows operating systems. While using this tool, user creates segments in audio file by playing the file and pressing a key each

time he hears a segment break. Segments are then transcribed by clicking on waveform and typing in an appropriate text. Transcriber allows using several channels for different speakers, thus not sharing Praat's overlapping speech limitation.

ANVIL (v4.5) [16] (v3.6) [17] is an annotation tool that works on Unix, Windows and Mac systems. It contains rich tier system and has an ability to specify relationships between different tiers [17]. User creates annotations by clicking at start and end points in the desired track, marking the annotation's interval. Additional information about the specific annotation is then entered by using menus or entering free text. Annotations can be added, deleted or redefined at any time [16].

ELAN (v2.4.1) [16] is a linguistic annotation tool used to annotate audio and video files. It allows grouping of annotations across multiple layers that are part of hierarchies. It works on Windows, Mac and Linux systems, and supports importing of transcriptions from other programs. Data can be annotated with use of both mouse and keyboard, and there are several export options allowing further data analysis.

EXMARaLDA (v1.3.2) [16] is a system whose objective is to provide a common framework by which the projects can share, exchange, reuse and store their multilingual data. It consists of a data model, set of XML formats and various tools for working with spoken language corpora. It uses a time based data model that is very similar to the data models found in Praat, ELAN and ANVIL, which makes data exchange between those tools a simple task. Most important tools are Partitur Editor, Corpus Manager and a Query Tool. These tools function on all major operating systems.

5 Overview of existing databases

There are few factors by which existing EMD can be classified by. Firstly, they can be classified by their scope, that includes number of subjects and number and range of emotions. Secondly, by the type of emotions - are they acted or natural? Thirdly, by the number of modes that is included in database, speech, face expressions, hand gestures, body pose, etc. And finally by the way emotional content is described [11].

Here is a short overview of some of the existing databases.

Belfast Natural Database [11] [14] is audio visual database containing video clips in English language taken from television chat shows, current affairs programmes and interviews conducted by research team. It contains totally 50 clips of naturalistic and induced material, containing a wide range of labeled emotional data.

Geneva Airport Lost Luggage Study [11] [14] database is an audio visual database containing unobtrusively videotaped clips of passengers at lost luggage counter followed up by interviews with passengers. Video clips are recorded in mixed languages and contain natural emotions of anger, humour, mixed indifference, stress and sadness.

Chung [11] [14] database is an audio visual database containing television interviews in English and Korean language in which speakers talk on a range of topics including sad and joyful moments in their lives. Naturalistic video clips contain emotions of joy, neutrality and sadness.

The HUMAINE Database [12] [14] is an audio visual database containing naturalistic and induced material in English, French and Hebrew languages. It contains a wide range of labeled emotional content including audio visual and gesture data. Content is labeled at 2 levels, globally across the whole emotion episode and time-aligned labeled continuously over time through the clip.

SMARTKOM Multimodal Corpus [13] [14] is an audio visual database containing interactive discourse with 224 speakers in German language. Emotional content includes joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise, and neutrality. Gesture data is also included.

SALAS Database [14] is an audio visual database in English language containing induced emotional content. Material is obtained from pilot study of 20 subjects talking to artificial listener while emotional states are changed by interaction with different personalities of the listener. It includes wide range of emotion related states but not very intense.

6 Conclusion and future work

This paper describes the creation of emotional multimodal databases and overviews their structure. They contain a collection of multimodal data that represents emotions and are the basis of emotionally intelligent systems. The objective of such systems is to improve the quality of human-machine interaction by improving the system's ability to recognize human emotion and reacting appropriately to user's mood changes.

Existing multimodal databases and database tools were overviewed and described.

In future studies we plan to construct a multimodal database that will contain speech in Croatian language and video for emotional interactions. We plan to start by recording limited number of speakers, annotate the data we obtain using the ELAN annotation tool, and afterwards keep upgrading the database as research continues.

References

[1] Hayamizu S., Hasegawa O., Itou K., Sakaue K., Tanaka K., Nagaya S., Nakazawa M., Endoh T., Togawa F., Sakamoto K., Yamamoto K.: RWC Multimodal Database for Interactions by Integration of Spoken Language and Visual Information, ICSLP 96 Proceedings, 3rd - 6th October, Philadelphia, USA, 1996, pp. 2171-2174.

[2] Picard, R.: Affective Computing, MA: MIT Press, Boston, USA, 1997.

[3] Frijda, N.H.: The Emotions, Cambridge University Press, Cambridge, UK, 1986

[4] Castellano G., Kessous L., Caridakis G.: Multimodal Emotion Recognition from Expressive Faces, Body Gestures and Speech, Proceedings of the Doctoral Consortium of the 2nd International Conference on Affective Computing and Intelligent Interaction, 13th - 14th September, Lisbon, 2007, pp. 375-388.

[5] Gajšek R., Štruc V., Mihelič F., Podlesek A., Komidar L., Sočan G., Bajec B.: Multi-Modal Emotional Database: AvID, Informatica, Ljubljana, Slovenia, 2009, pp. 101-106.

[6] Le Chenadec G., Maffiolo V., Chateau N., Colletta J.: Creation of a Corpus of Multimodal Spontaneous Expressions of Emotions in Human-Machine Interaction, Proceedings of 5th International Conference on Language Resources and Evaluation, 24th - 26th May, Genova, Italy, 2006.

[7] Chitu A.G., van Vulpen V., Takapoui P., Rothkrantz L.J.M.: Building a Dutch Multimodal Corpus for Emotion Recognition, LREC 2008 Workshop on Corpora for Research on Emotion and Affect, May, ELRA, ELRA, 2008, pp. 53-56.

[8] Martin J.C., Abrilian S., Devillers L.: Annotating Multimodal Behaviors Occurring During non Basic Emotions, Affective Computing and Intelligent Interaction Proceedings, 22nd - 24th October, Beijing, China, 2005, pp. 550-557.

[9] Noguchi M., Miyoshi K., Tokunaga T., Iida R., Komachi M., Inui K.: Multiple Purpose Annotation using SLAT - Segment and Link-based Annotation Tool -, Proceedings of 2nd Linguistic Annotation Workshop, 26th - 27th May, M, M, 2008, pp. 61-64.

[10] Rohlfing K., Loehr D., Duncan S., Brown A., Franklin A., Kimbara I., Milde J.T., Parrill F., Rose T., Schmidt T., Sloetjes H., Thies A., Wellinghoff S.: Comparison of Multimodal Annotation Tools - Workshop Report, Gesprachforschung - Online-Zeitschrift zur Verbalen Interaktion, vol. 7, 2006, pp. 99-123.

[11] Douglas-Cowie E., Campbell N., Cowie R., Roach P.: Emotional Speech: Towards a New Generation of Databases, Speech Communication, vol. 40, 2003, pp. 33-60.

[12] Douglas-Cowie E., Cowie R., Sneddon I., Cox C., Lowry O., McRorie M., Martin J.C., Devillers L., Abrilian S., Batliner A., Amir N., Karpouzis K.: The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data, Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, 13th - 14th September, Lisbon, 2007, pp. 488-500.

[13] Schiel F., Steininger S., Turk U.: The SmartKom Multimodal Corpus at BAS, Proceedings of the 3rd Language Resources and

Evaluation Conference, 29th - 31st May, Las Palmas, Gran Canaria, Spain, 2002, pp. 200-206.

[14] The HUMAINE Association: Databases, available at `http://emotion-research.net/wiki/Databases`, Accessed: 1st May 2010.

[15] Zimmermann P., Guttormsen S., Danuser B., Gomez P.: Affective Computing - A Rationale for Measuring Mood with Mouse and Keyboard, International Journal of Occupational Safety and Ergonomics, vol. 9, Warsaw, Poland, 2003, pp. 539-551.

[16] Rohlfing K., Loehr D., Duncan S., Brown A., Franklin A., Kimbara I., Milde J.T., Parrill F., Rose T., Schmidt T., Sloetjes H., Thies A., Wellinghoff S.: Comparison of Multimodal Annotation Tools - Workshop Report, Gesprachforschung - Online-Zeitschrift zur Verbalen Interaktion, vol. 7, 2006, pp. 99-123.

[17] Garg S., Martinovski B., Robinson S., Stephan J., Tetreault J., Traum D.R.: Evaluation of Transcription and Annotation Tools for a Multi-Modal, Multi-Party Dialogue Corpus, Proceedings of 4th Language Resources and Evaluation Conference, 26th - 28th May, Centro Cultural de Belem, Lisbon, Portugal, 2004, pp. 2163-2166.

[18] Horvat M., Popović S., Bogunović N., Ćosić K.: Tagging Multimedia Stimuli with Ontologies, Proceedings of the 32nd International Convention MIPRO, 25th - 29th May, Rijeka, Croatia, 2009, pp. 203-208.

[19] Ćosić K., Popović S., Kukolja D., Horvat M., Dropuljić M.: Physiology-Driven Adaptive Virtual Reality Stimulation for Prevention and Treatment of Stress Related Disorders, Cyberpsychology, Behavior, and Social Networking, vol. 13, 2010, pp. 73-78.

[20] Peter C., Herbon A.: Emotion representation and physiology assignments in digital systems, Interacting with Computers, vol. 18, New York, USA, 2006, pp. 139-170.