Building a Dutch Multimodal Corpus for Emotion Recognition

Alin G. Chițu, Mathijs van Vulpen, Pegah Takapoui and Leon J.M. Rothkrantz
Faculty of Information Technology and Systems
Delft University of Technology
Mekelweg 4, 2628CD Delft,
The Netherlands
E-mails: {A.G.Chitu,L.J.M.Rothkrantz}@ewi.tudelft.nl, mathijs@ch.tudelft.nl, pegahtak@gmail.com

Abstract

Multimodal emotion recognition gets increasingly more attention from the scientific society. Fusing together information coming on different channels of communication, while taking into account the context seems the right thing to do. During social interaction the affective load of the interlocutors plays a major role. In the current paper we present a detailed analysis of the process of building an advanced multimodal data corpus for affective state recognition and related domains. This data corpus contains synchronized dual view acquired using high speed camera and high quality audio devices. We paid careful attention to the emotional content of the corpus in all aspects such as language content and facial expressions. For recordings we implemented a TV prompter like software which controlled the recording devices and instructed the actors to assure the uniformity of the recordings. In this way we achieved a high quality controlled emotional data corpus.

1. Introduction

The affective state of a person is very important in human communication. During social interaction humans express their affective state through a large variety of channels, such as facial expressions, communicative gestures like body posture, emotional speech, etc. The semantic content of our communication is largely enriched by transmitting to the interlocutor our current affective state. The affective state influences the way we interact with our interlocutors, our actions and reactions to certain situations. Also, in the case of human computer interaction, it would greatly increase the quality of our experiences if the machine would be able to adapt to our affective state. We can imagine for instance that we are involved into a crisis situation and we use our PDA to communicate to and receive indications from a central crisis management center. Knowing the affective state of the user the system can adapt the content and layout of the messages to increase their receptivity. The system can do this transparently, for all users without requiring that the sender is aware of this. In this way we can optimize the search and rescue activities. There are many other applications of affective state recognition, to name a few more: children toys which can tailor to the children needs in each moment, any public kiosks, ATMs, driver safety systems, etc.

As in the case of speech recognition (McGurk and MacDonald 1976) people use context information acquired through different communication channels to improve the accuracy of the affective state recognition. For instance speech and emotion recognition are two much interconnected processes, which influence each other. The exact influence is not completely elucidated. Our speech influences the facial expressions and our facial expressions influence our speech. Of course the affective state of the speaker is largely transmitted through prosody. Buchan et. al. (Buchan et. al. 2007) analyzed what the subjects are watching while trying to understand what people are saying or what facial expressions are they showing. They showed that the distribution of gaze is dependent on the distribution of information in the face and on the goals of the user. It was concluded as well that emotion related information is spread on the entire face. Notable is for instance the concentration of the gaze around the nose when the signal to noise ratio decreases.

Data corpora are an important building block of any scientific study. The data corpus should provide the means for understanding all the aspects of a given process, direct the development of the techniques toward an optimum solution by allowing for the necessary calibration and tuning of the methods and also give good means for evaluation and comparison. Having a good data corpus (i.e. well designed, capturing both general and also particular aspects of a certain process) is of great help for the researchers in this field as it greatly influences the research results. Having this in mind we decided to build such a data corpus. A good data corpus should have a good coverage of the process it going to be investigated such that every aspect should get a fair slice.

We present in this paper a detailed analysis of the process of building an advanced multimodal emotion data corpus for the Dutch language. We strongly believe that sharing our experiences is the first step for understanding the issues around building a reliable data corpus. We envision a future standard for data corpora that combines the views of the entire scientific community.

2. Recordings' settings

This section presents the settings used while compiling the data corpus. Figure 1 shows the complete image of the setup. We used a high speed camera, a professional microphone and a mirror for dual view synchronization. The camera was controlled by the speaker, through a prompter like software. The software was presenting the speaker the next item to be uttered together with directions on the speaking style required. This provided us with a better control of the recordings.

2.1 Audio and Video devices

The audio and video quality is an important issue to be covered. An open question is for instance, what is the optimum sampling rate in the visual domain? Current standard for video recording frame rate ranges from 24 up to 30 frames per second, but is that enough? A first problem and the most intuitive is the difficulty in handling the increased amount of data, since the bandwidth needed is many times larger. A second problem is a technical problem and is related with the techniques used for fusing the audio and video channels. Since it is common practice to sample the audio stream at a rate of 100 feature vectors per second, in the case when the information is fused in an early stage, we encounter the need to use interpolation to match the two data sampling rates. A third issue, that actually convinced us to use a high speed camera, is related to the coverage of the visemes during recording, namely the number of frames per visemes. In the paper Chițu and Rothkrantz 2007 it was showed that the visemes coverage becomes a big issue when the speech rate increases. While talking with experts from the brain and speech domain we learned that recording at 125Hz should cover almost every movement on a person's face. There are, however, movements like the lips vibration when the air is pushed with high speed through the loosely closed lips that require some 400Hz for exact recording. Therefore we decided to use a high speed camera for video recordings. As we aim to discover where the most useful information for emotion detection lies and we want to give the possibility for developing new applications we decided to include side view recordings of the speaker's face in our corpus.

When one goes outside the range of consumer devices, things become extremely more complicated and definitely more expensive. The quality of the sensors and the huge bandwidth necessary to stream high speed video to the PC makes high speed video recording very restrictive. We used for recording a Pike F032C camera built by AVT. The camera is capable of recording at 200Hz in black and white, 139Hz when using the chroma subsampling ratio 4:1:1 and 105Hz when using the chroma subsampling ratio 4:2:2 while capturing at maximum resolution 640X480. By setting a lower ROI the frame rate can be increased. In order to increase the Field Of View (FOV), as we will mention later, we recorded in full VGA resolution. To be able to guarantee a fix and uniform sampling rate and to permit an accurate synchronization with the audio signal we used a pulse generator as an external trigger. A sample frame is shown in Figure 2. To acquire a synchronized dual view we used a mirror which was placed behind the speaker at 45° (see Figure 1).

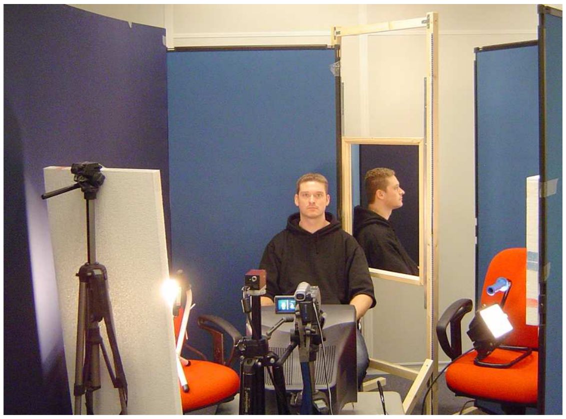


Figure 1: The setup of the experiment.

In the case of video data recording there are a larger number of important factors that control the success of the resulted data corpus. Hence, not only the environment, but also the equipment used for recording and other settings is actively influencing the final result. The environment where the recordings are made is very important since it can determine the illumination of the scene, and the background of the speakers. We use mono-chrome background so that by using a "chroma keying" technique the speaker can be placed in different locations inducing in this way some degree of visual noise.

For recording the audio signal we used NT2A Studio Condensators. We recorded a stereo signal using a sample rate of 48kHz and a sample size of 16bits. The data was stored in PCM audio format. The recordings were conducted in controlled laboratory environment. We considered that it is more advantageous to have very good quality recordings and degrade them in a post process as needed. The specific noise can be simulated or recorded in the required conditions and later superimposed on the clear audio data. An example of such database is NOISEX-92 (Varga and Steeneken 1993). This dataset contains white noise, pink noise, speech babble, factory noise, car interior noise, etc.

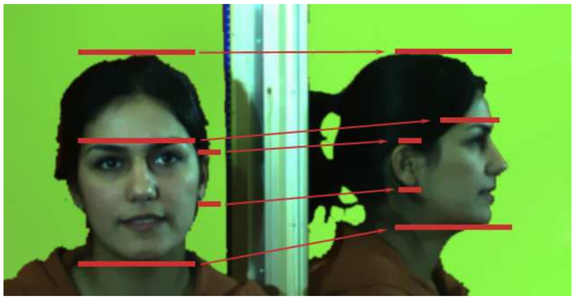
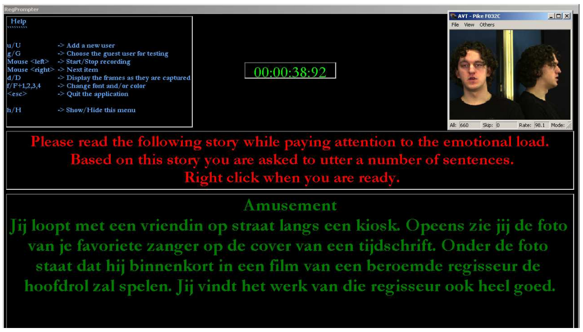


Figure 2: Sample frame with dual view.

2.2 The Prompter Tool

Using a high speed camera increases the storage needs for the recordings. It is almost impossible to record everything and than during the annotation process, cut the clips at the required lengths. One main reason is that when recording in high speed high resolution the bandwidth limitation requires that the video be captured in the memory (e.g. on a RAM Drive). This makes the clips to have a maximum length of approximately 1 minute, depending on the resolution and color subsampling ratio used. However, we needed anyway to present the speakers with the pool of items required to be uttered. We build therefore a prompter like tool that provided the user the next item to be uttered together with some instructions about the speaking style and also controlled the video and audio devices. The result was synchronized audio and video clips already cropped to the exact length of the utterance. The tool provided the speaker the possibility to change the visual themes to maximize the visibility, and offer a better recording experience.



```
u/U             -> Add a new user
g/G             -> Choose the guest user for testing
Mouse <left>    -> Start/Stop recording
Mouse <right>   -> Next item
d/D             -> Display the frames as they are captured
f/F+1,2,3,4     -> Change font and/or color
<esc>           -> Quit the application

h/H             -> Show/Hide this menu
```

Please read the following story while paying attention to the emotional load.
Based on this story you are asked to utter a number of sentences.
Right click when you are ready.

Amusement
Jij loopt met een vriendin op straat langs een kiosk. Opeens zie jij de foto van je favoriete zanger op de cover van een tijdschrift. Onder de foto staat dat hij binnenkort in een film van een beroemde regisseur de hoofdrol zal spelen. Jij vindt het werk van die regisseur ook heel goed.

Figure 3: Prompter view during recordings.

The control of the software was done by the speaker through the mouse buttons of a wireless mouse that was taped on the arm of the chair. After a series of trials we conclude that this level of control is sufficient and not very disruptive for the speaker. The tool was also used to keep track of the user's data, recording takes and recording sessions.

2.3 Emotional speech

There are two different approaches to collect data for an emotion database: by capturing real data or by inducing the emotional status to the actors. The first approach is almost impossible to be used because of all the ethical issues linked with trust and personal intimacy. Therefore we collected a set of stories which carried a strong emotional load. We asked each speaker to read each story and then transpose him/herself into the right affective state and utter a set of 5 appropriate sentences as a possible reaction to the particular story. Of course a good question regarding this approach would be whether the quality of the expressed emotions is preserved, or the recorded material contains artificial performances. In real life it is very difficult to select isolated emotions; usually people show an amalgam of emotions. The speakers were divided into two groups: professional actors and naive speakers. All speakers were native Dutch. This is very important for the case of emotional speech since the performance of the speaker could get less genuine and definitely less spontaneous as result of the speaker spending more time in preparing his speech. However, it could be very interesting to analyze the cultural effect on expressing ones' emotions through facial expressions and prosody. We recorded 21 emotions which are listed in Table 1. An example of the story and reactions used for recordings is given in Table 2.

#	Emotion	#	Emotion
1	Admiration	**12**	Fear
2	Amusement	**13**	Fury
3	Anger	**14**	Happiness
4	Boredom	**15**	Indignation
5	Contempt	**16**	Interest
6	Desire	**17**	Pleasant surprise
7	Disappointment	**18**	Unpleasant surprise
8	Disgust	**19**	Satisfaction
9	Dislike	**20**	Sadness
10	Dissatisfaction	**21**	Inspiration
11	Fascination		

Table 1: List of emotions considered for recordings.

Dutch original
Emotie: "Bewondering" **Vertelling:** "Je loopt samen met een vriend/vriendin door een dure winkelstraat in Amsterdam en ziet in de etalage een jas hangen die je altijd al had willen hebben. Je droomt over wat je zou doen als je het geld had om deze jas te kopen. Je gaat voor de etalage staat en denkt..." **Reactie:** R1: Oooohhh... R2: Dat ziet er goed uit! R3: Die zou ik graag hebben! R4: Was die maar van mij! R5: Zodra ik mijn geld heb, is die jas van mij!
English approximative translation
Emotion: "Admiration" **Story:** "You walk together with your friend/girlfriend in front of a fancy store in Amsterdam and you see in the store's window a coat that you always wanted. You dream of what you would do if you have had the money to buy the coat. You stand in front of the window and think.... **Reaction:** R1: Oooohhh... R2: That looks so nice! R3: I would really want it! R4: That is for me! R5: As soon as I'll have money, that coat is mine.

Table 2: Story and possible reactions for "admiration".

3. Demographic data recorded

As we specified in the introduction a proper coverage of the variability of the speakers is needed to assure the success of a data corpus. We also have seen that there is a language use difference between speakers. This can be used for instance to develop adaptive recognizers. Therefore we recorded for each speaker the following data: gender, age, education level, native language (as well as whether he/she is bi-lingual) and region where he/she had grown up. The last aspect is used to identify possible particular clusters in the pool of actors. The cultural background of the actors can play an important role in the expressions showed. Persons from different cultures might give different meaning to different gestures and expressions. In our case since we only collect data based on native Dutch speakers we expect that the cultural impact to be reduced. However, it is a matter that should be investigated anyway.

4. Research goals and usability of the resulted data corpus

As we specified in the introduction the presented corpus targets the domain of multimodal affective state recognition. However, we have a large interest in analyzing the degree in which the emotional content and the speech content interfere. Hence we would like to be able to describe the impact of the affective state on the visemes shown by the speaker.

We also envision that by analyzing the data recorded we will be able to develop a formal way for annotating and describing such affective data.

We also expect that the resulted data corpus will enable the analysis of the recording quality, especially of the video sampling rate on the recognition results.

5. Data corpus size

The duration of each recording session was approximately 45 minutes. Each session resulted in a number of 105 performances recorded by the actor. Hence each actor recoded approximately 15 minutes. We collected data from 25 persons, mainly students at our technical university (of course we also took advantage of the rest of the stuff in our department). We would like however that our complete data corpus to contain data from at least 50 actors. We also have access to a number of professional actors which agreed to take part in our experiment. This set is particularly important because their performances are going to be used for assessing the quality of the acted emotions by the rest of the actors. Hence in total we expect to collect more than 5000 performances.

6. Conclusions

We presented in this paper our thoughts and investigations on building a good data corpus. We presented the settings used during the recordings, the language content and the recordings progression. The new data corpus should consist of high speed recordings of synchronized dual view of speaker faces while uttering emotional speech and showing the appropriate facial expressions. It should provide a sound tool for training, testing, comparison and tuning a highly accurate affective state recognizer. There are still many questions to be answered with respect to building a data corpus. For instance which modalities are important for a given process, and moreover what is the relationship between these modalities. Is there any important influence between different modalities?

A major issue to be addressed is the quality of the acted data. As we specified we plan to use the recordings of the professional actors to assess the quality of the rest of the naïve actors.

Our data corpus only contain recordings with individuals showing emotions triggered by reading some emotional stories, however we only consider scenes with single actors showing “clean” emotions. However, it has been shown that there are multiple situations in real life when people show in fact an amalgam of emotions. This issue should be address as well.

7. Acknowledgements

The work reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

8. References

(Buchan et. al. 2007) Julie N. Buchan, Martin Paré and Kevin G. Munhall, „Spatial statistics of gaze fixations during dynamic face processing”, Journal of Social Neuroscience, 2007, vol 2, 1-13.

(Chițu and Rothkrantz 2007) Alin G. Chițu and Leon J.M. Rothkrantz, "The Influence of Video Sampling Rate on Lipreading Performance", *12-th International Conference on Speech and Computer (SPECOM'2007)*, ISBN 6-7452-0110-x, pp. 678-684, Moscow State Linguistic University, Moscow, October 2007.

(McGurk and MacDonald 1976) McGurk, H. & MacDonald, J. Hearing lips and seeing voices *Nature*, 1976, *264*, 746 – 748.

(Varga and Steeneken 1993) Varga, A. and Steeneken, H. 1993. “Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems.” *Speech Communication*, (vol. 12, no. 3, pp. 247-251, July).