Programme of the Workshop on Corpora for Research on Emotion and Affect

9.00 Welcome

Induced emotions (chairman A. Batliner)

9.05 The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation. *Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amir, Dirk Heylen*

9.25 The NIMITEK Corpus of Affected Behavior in Human-Machine Interaction: *Milan Gnjatovic and Dietmar Roesner*

9.55 An Interface to Simplify Annotation of Emotional Behaviour *Shazia Afzal, Peter Robinson*

10.05 - 10.45 Coffee break (pose posters)

Acted versus spontaneous emotions (chairman R. Cowie)

10.45 Anger detection performances based on prosodic and acoustic cues in several corpora *Laurence Vidrascu, Laurence Devillers, LIMSI-CNRS, France*

11.05 Recording audio-visual emotional databases from actors: a closer look: *Carlos Busso and Shrikanth S. Narayanan*

11.25 Acted vs. spontaneous expressive speech: perception with inter-individual variability: *Nicolas Audibert, Véronique Aubergé and Albert Rilliard*

Realistic emotions corpora (chairman E. Cowie)

11.45 Releasing a thoroughly annotated and processed spontaneous emotional data: the FAU Aibo Emotion Corpus; *A. Batliner, S. Steidl, E. Nöth*

12.05 Emotional Speech Corpus Construction, Annotation and Distribution, *Charlie Cullen, Brian Vaughan, Spyros Kousidis, John McAuley*

12.25 Emotions in a Corpus of Human and Computer Keyboard-to-Keyboard Tutoring Sessions: *Farhana Shah, Martha Evens*

12h45 - Discussion

13.00 - 14.30 Lunch

14.30 Vocal expression in spontaneous and experimentally induced affective speech: Acoustic correlates of anxiety, irritation and resignation, *Petri Laukka, Kjell Elenius, Mats Fredrikson, Tomas Furmark and Daniel Neiberg*

SESSION "POSTER"

Emotional corpora (chairman L. Devillers)

15h – 17h

1- Multimodal records of driving influenced by induced emotion
Edelle McNahon, Roddy Cowie, Johannes Wagner, Elisabeth André

2- Building a Dutch Multimodal Corpus for Emotion Recognition
Alin G. ChiŃu, Mathijs van Vulpen, Pegah Takapoui and Leon J.M. Rothkrantz

3- Multi-modal emotion-related data collection within a virtual earthquake emulator

Dimitrios Ververidis, Irene Kotsia, Constantine Kotropoulos, and Ioannis Pitas

4- TRUE: an Online Testing Platform for Multimedia Evaluation
Santiago Planet, Ignasi Iriondo, Elisa Martínez, José A. Montero

5- Spanish Expressive Voices: corpus for emotion research in Spanish
Roberto Barra-Chicote, Juan Manuel Montero, Javier Macias-Guarasa, Syaheerah Lufti, Juan Manuel Lucas, *Fernando Fernandez, Luis Fernando D'haro, Ruben San Segundo, Javier Ferreiros, Ricardo Cordoba and Jose Manuel Pardo*

6- IrcamCorpusExpressivity: Nonverbal Words and Restructurings
Grégory Beller, Christophe Veaux and Xavier Rodet

7- Dynamic Detection of Mood Propagation in Fan Groups
Gail L, Dave Bruckmayr; Paulo Barthelmess

8- Ambiguous classification of emotional speech

Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou and Liming Chen

16.00-16.30 Coffee break

Emotional Multimodal analysis and interaction (chairman JC. Martin)

16.30 Testimonials on emotions – a multimodal speech analysis
Gaëlle Ferré

16.50 Static vs. dynamic Gestural Icons of "Feeling of Thinking", Vanpé Anne, Aubergé Véronique, GIPSA Lab - Institut de la Communication Parlée, Grenoble.

17.10 Cooperation Attitude in Negotiation Dialogs : *.Nicole Novielli, University of Bari, Italy, Peter Carnevale, Jonathan Gratch*

17.30 Developing Affective Intelligence For An Interactive Installation: Insights From A Design Process, *Lassi A. Liikkanen, Eero Huvio, Rodolfo Samperio, Eero Väyrynen*

18.00-19.00 Discussion (R. Cowie)
19.00 End of workshop (followed by an informal dinner)

Organiser(s)

DEVILLERS, L. (CNRS-LIMSI, France)

MARTIN, J.-C. (CNRS-LIMSI, France)

COWIE, R. (QUB : Queen Nelfast Univ, Irland)

DOUGLAS-COWIE E. (QUB : Queen Nelfast Univ., Irland)

BATLINER, A. (Erlangen Univ, Germany.)

Programme Committee

John Hansen
Elisabeth Schriberg
Marc Schroeder
Shrikanth Narayanan
Kostas Karpouzis
Ioana Vasilescu
Noam Amir
Nick Campbell
Fiorella de Rosis
Isabella Poggi
Nadia Berthouze

Table of Contents

Introduction

This decade has seen an upsurge of interest in systems that register emotion (in a broad sense) and react appropriately to it. Emotion corpora are fundamental both to developing sound conceptual analyses and to training these 'emotion-oriented systems' at all levels - to recognize user emotion, to express appropriate emotions, to anticipate how a user in one state might respond to a possible kind of reaction from the machine, etc. Corpora have only begun to grow with the area, and much work is needed before they provide a sound foundation.

This workshop follows a first successful workshop on Corpora for Research on Emotion and Affect at LREC 2006. Papers are in the area of corpora for research on emotion and affect. They raise one or more of the following questions. What kind of theory of emotion is needed to guide the area? What are appropriate sources? Which modalities should be considered, in which combinations? What are the realistic constraints on recording quality? How can the emotional content of episodes be described within a corpus? Which emotion-related features should a corpus describe, and how? How should access to corpora be provided? What level of standardization is appropriate? How can quality be assessed? What are the ethical issues in database development and access?

The organizers are members of the Humaine association (http://emotion-research.net).

The organizing committee:

Laurence Devillers / Jean-Claude Martin
Spoken Language Processing group/ Architectures and Models for Interaction,
LIMSI-CNRS,
BP 133, 91403 Orsay Cedex, France
(+33) 1 69 85 80 62 / (+33) 1 69 85 81 04 (phone)
(+33) 1 69 85 80 88 / (+33) 1 69 85 80 88 (fax)
devil@limsi.fr / martin@limsi.fr
http://www.limsi.fr/Individu/devil/
http://www.limsi.fr/Individu/martin/

Roddy Cowie / School of Psychology
Ellen Douglas-Cowie / Dean of Arts, Humanities and Social Sciences
Queen's University, Belfast BT7 1NN, UK
+44 2890 974354 / +44 2890 975348 (phone)
+44 2890 664144 / +44 2890 ****** (fax)
http://www.psych.qub.ac.uk/staff/teaching/cowie/index.aspx
http://www.qub.ac.uk/en/staff/douglas-cowie/
r.cowie@qub.ac.uk / e.douglas-Cowie@qub.ac.uk

Anton Batliner - Lehrstuhl fuer Mustererkennung (Informatik 5)
Universitaet Erlangen-Nuernberg - Martensstrasse 3
91058 Erlangen - F.R. of Germany
Tel.: +49 9131 85 27823 - Fax.: +49 9131 303811
batliner@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de/Personen/batliner/

The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation

Ellen Douglas-Cowie[1], Roddy Cowie[1], Cate Cox[1], Noam Amir [2], Dirk Heylen [3]
Queen's University Belfast [1], Tel Aviv University [2], University of Twente [3]
e-mail: e.douglas-cowie@qub.ac.uk, r.cowie@qub.ac.uk, c.cox@qub.ac.uk, noama@post.tau.ac.il, heylen@cs.utwente.nl

Abstract

The aim of the paper is to document and share an induction technique (The Sensitive Artificial Listener) that generates data that can be both tractable and reasonably naturalistic. The technique focuses on conversation between a human and an agent that either is or appears to be a machine. It is designed to capture a broad spectrum of emotional states, expressed in 'emotionally coloured discourse' of the type likely to be displayed in everyday conversation. The technique is based on the observation that it is possible for two people to have a conversation in which one pays little or no attention to the meaning of what the other says, and chooses responses on the basis of superficial cues. In SAL, system responses take the form of a repertoire of stock phrases keyed to the emotional colouring of what the user says. The technique has been used to collect data of sufficient quantity and quality to train machine recognition systems.

1 Introduction

It is a difficult problem to generate recordings of emotionally coloured conversation data that are reasonably natural, but still suitable for machine analysis. This paper describes an induction method that generates data which has been successfully used in a machine learning environment. The technique is called the Sensitive Artificial Listener Technique, developed at Queen's University Belfast. The aim is to document this tool and share it with the research community.

There have been several published descriptions of analyses that use data from SAL exercises, but there is no generally available description of the technique itself. This paper remedies that omission.

1.1 Background and Context

It has become clear that for different reasons, emotion-oriented computing cannot rely either on data from actors or on fully naturalistic recordings. As a result, there is great interest in data generated by techniques designed to elicit emotion deliberately. This type of approach produces data that can be both tractable and reasonably naturalistic. Many induction techniques are in use in the machine learning context, such as computer games (Bechara, Damasio, Damasio & Anderson 1994; van Reekum et al 2004; Wang and Marsella 2006) or tasks involving computers (Batliner, Fischer et al, 2003; Batliner, Hacker et al. 2003; Aubergé, Audibert & Rilliard 2004) and sometimes tasks involving human-human interaction (Bachorowski & Owren 1995; Abassi et al 2007; Martin et al. 2006).

The Sensitive Artificial Listener is a specific type of induction technique that focuses on conversation between a human and an agent that either is or appears to be a machine. It is designed to capture a broad spectrum of emotional states, expressed in 'emotionally coloured discourse' of the type likely to be displayed in everyday conversation.

It is a challenge to collect records of human-machine conversation, because machines are not actually able to carry out conversations. However, there are obvious reasons to try, since it seems very likely that human-machine interactions will differ from human-human interactions in significant ways. Not the least of these is that for the foreseeable future, human-machine interactions will break down in ways that human-human interactions do not, and it is important to have ways of recognising the signs of breakdown.

2 The SAL technique

2.1 The basic context

The Sensitive Artifical Listener technique (SAL for short) is based on the observation that it is possible for two people to have a conversation in which one pays little or no attention to the meaning of what the other says, and chooses responses on the basis of superficial cues. The point was made long ago by the ELIZA scenario (Weizenbaum 1996). In the SAL technique, system responses are keyed to the emotional colouring of what the user says, rather than (as in Eliza) words or phrases. The versions used so far have used Wizard of Oz techniques where a human operator follows a script that specifies possible responses. Because the aim is to evoke emotionally coloured responses, the statements are stock phrases chosen to evoke strong reactions in the listener. In current versions, the SAL operator chooses which statement to use at any given time from a menu that is organised to simulate four personalities – Poppy (who aims to make people happy), Obadiah (who aims to make people gloomy), Spike (who aims to make people angry) and Prudence (who aims to make people pragmatic). Users choose at any time which 'personality' they want to talk to. The response that is chosen will depend on the 'personality' that is active and the user's state. The combination creates an environment rich enough to provoke exchanges that are extended, and quite highly coloured emotionally.

2.2 The SAL structure

The four characters are equipped with a set of characteristic responses encouraging the user into responding in differing emotional states. The SAL has no intelligence, only prespecified stock responses.

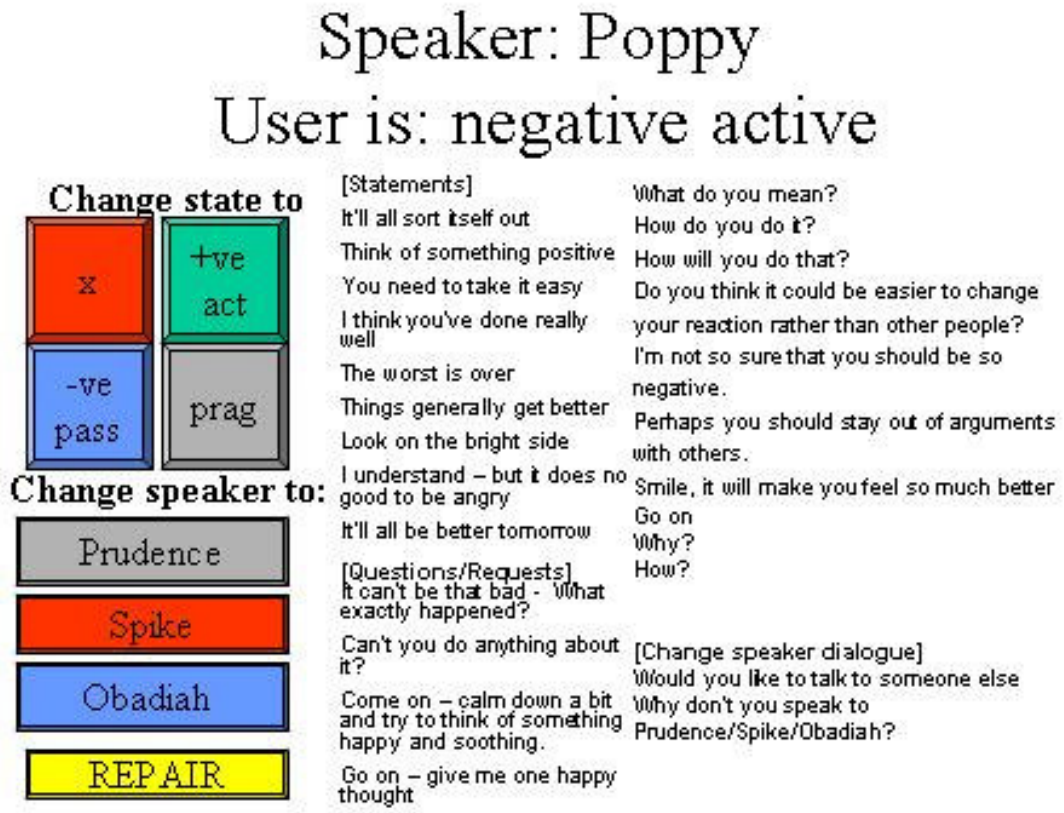


FIGURE 1: Structure governing interaction in SAL

The scripts for the characters were developed, tested and refined in an iterative way. Each character has a number of different types of script depending on the emotional state of the user. So, for example, Poppy has a script for the user in each of four emotional states - a positive active state, a negative active state, a pragmatic state, a negative passive state. There are also script types relevant to the part of the conversation (beginning, main part) or structural state of the conversation (repair script).Each script type has within it a range of statements and questions. They cannot be context specific as there is no 'intelligence' involved. Figure 1 illustrates the type of structure that governs an interaction between one of the personalities of SAL (taken from an early version of SAL).

There have been different versions of SAL moving from an early Wizard of Oz version where the scripts for each personality/character are read by an experimenter who used different tones of voice for the four characters (SAL 0) to a more sophisticated version developed in conjunction with the University of Twente where the phrases are pre-recorded and the experimenter selects phrases from a menu (SAL 1). A fully automated version is currently being developed under the SEMAINE project (http://www.semaine-project.eu/).
The original version of SAL was in English and was successful enough for versions to be developed in Hebrew (at Tel Aviv University) and Greek (at National Technical University of Athens, ICCS) with adjustments to suit cultural norms and expectations.

2.2 User experience

SAL has been described as an emotional gym. It does not manipulate users' emotions: that would need far more sophistication. Rather, it gives them prompts to which they can react emotionally if they choose to. It is easy to build up quite a high level of involvement during a sequence of exchanges on an emotive topic. That may be partly because SAL does not inhibit emotional expression by introducing different subjects or perspectives. Various factors lead engagement to break down eventually. SAL responses may simply be too ridiculous for the user to accept; they may become too repetitious; the user may become hopelessly frustrated with SAL's inability to answer questions. Nevertheless, experienced users in particular can easily sustain quite protracted conversations with the system, on the order of half an hour. It appears that listeners learn to use the system, which means that longitudinal use by small numbers has some advantages over occasional use by many.

3 Data

The SAL scenario has been used successfully in three major EU projects (ERMIS, HUMAINE and SEMAINE) to generate large amounts of data that has been labelled and used in a machine learning context.

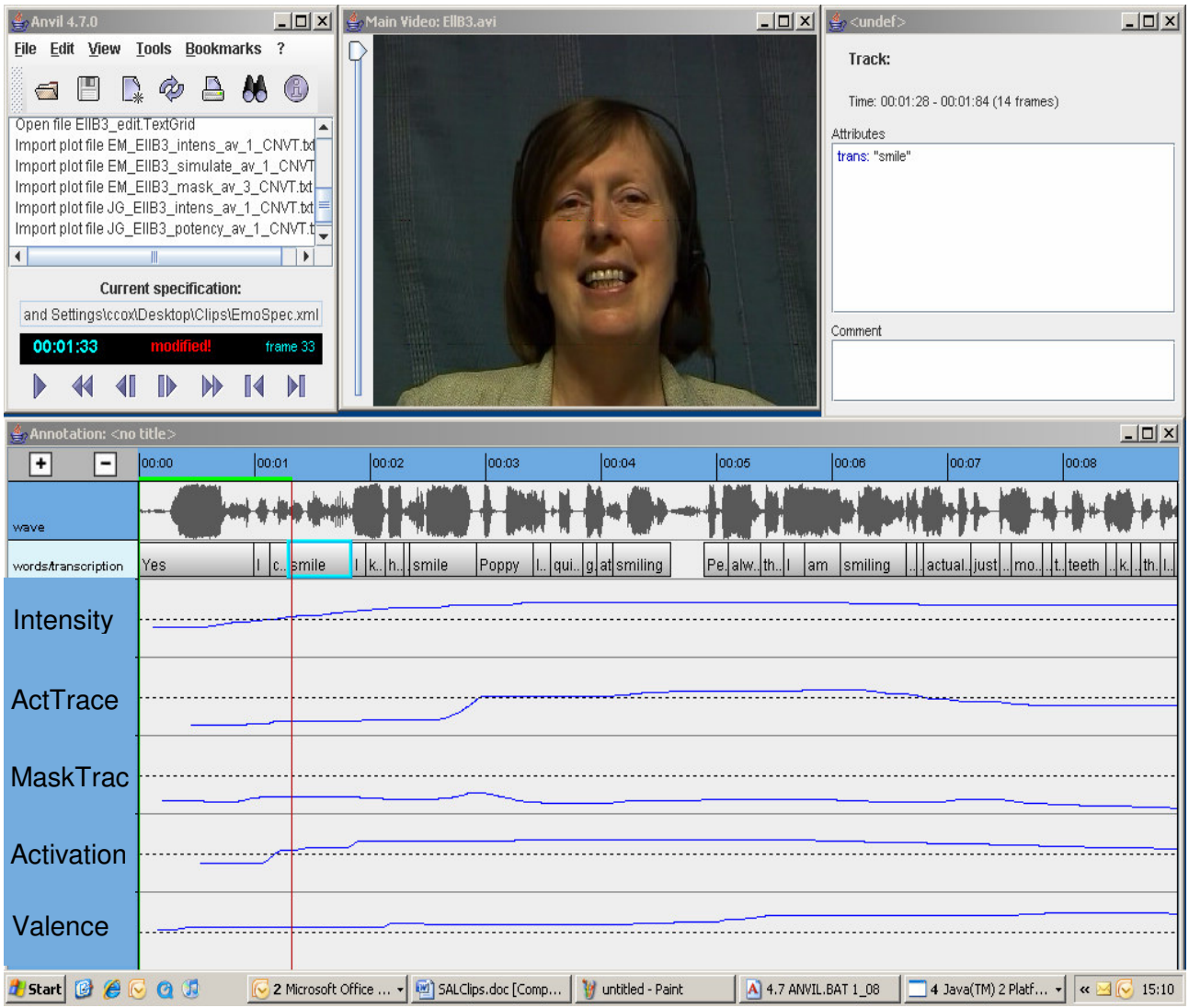


Figure 2: Labelled SAL data from the HUMAINE Database

The data generated is rich in facial and non verbal signals (e.g. aspects of pitch, spectral characteristics, timing), and shows a considerable range of emotions and emotional intensities.

Data was collected using the SAL 0 version from 20 users, 10 male and 10 female. In total 105 minutes of footage was collected . This has been segmented into files and is in avi and mpeg format. There are accompanying files of what was said. SAL 1 was used to collect data from four users, each recorded for two sessions, each of approximately 30 minutes. The data is segmented into files in avi and mpeg format and four raters have labelled the data using the dimensional FEELtrace tool (see Cowie et al. 2000). This gives labels on two dimensions related to emotion (activation and evaluation), and produces traces of how a user's emotional state is perceived over time. A substantial body of data from SAL 1 has also been labelled in a more detailed way as part of the HUMAINE Database (www.emotion-research.net/download/pilot-db/). Data from SAL 1 is releasable under an agreement governing the use of the data. The Hebrew version has undergone a number of translations and data has now been collected from 5 users, totalling 2.5 hours.

Figure 2 illustrates the kind of data that is produced. It presents a SAL 1 sequence labelled as part of the HUMAINE Database. The emotionality in the user's facial expression is evident. This is borne out by the accompanying traces from a rater for emotion intensity and activation – first and fourth trace lines respectively. The point at which the shot of the face is taken (marked by the vertical red line) corresponds to a rise in the emotional intensity and degree of activation perceived by the rater. The second and third trace lines respectively indicate the degree to which the rater perceives the user to be acting or masking her emotion. The pattern of the ActTrace line indicates a low level of perceived acting at the start rising to absence of acting as the intensity of the emotion rises, indicating the naturalness of the emotion generated.

It is beyond the scope of this paper to describe the statistical properties of the ratings, but Figure 3 summarises some key points. It shows ratings of

valence (x axis) and activation (y axis) in the circular FEELtrace space from two raters. The data covers most of the space except strong negative emotion. The raters agree on the broad pattern, but one (data points in white) is more conservative. Ensuring acceptable consistency is too complex an issue to address here.

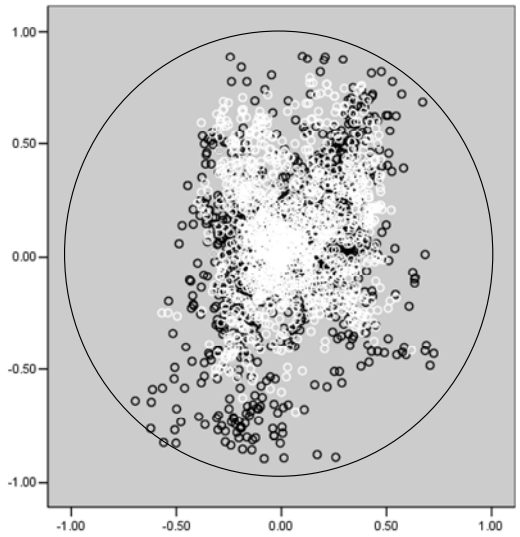


Figure 3: Emotional spread of SAL 1 data

The SAL data that is already available is of sufficient quantity and quality to train machine recognition systems. Published reports of research using the material include Ioannou et al. (2005) and Fragopanagos & Taylor (2005). More recent research reports very high recognition rates when the multimodal character of the data is exploited (Kollias et al. 2008).

4 Conclusion

The success of SAL has led to a new EU funded project called SEMAINE (http://www.semaine-project.eu/) which aims to build an automatic human-computer conversation system based on SAL. It will identify the user's emotional state itself, using evidence from face, upper body, voice, and key words. Its range of replies will include some ELIZA-like use of key words extracted from the user's speech. Its own speech will be synthesised, not recorded, and express its emotional stance towards the user. That stance will also be expressed through a graphical display of the 'listener's' face and shoulders. While the user is speaking, the 'listener' will also use vocalisations, facial expressions, and gestures (e.g. nodding) to signal its stance and prompt the speaker to continue or break.

The point of the project is that SAL provides a context in which sustained emotionally coloured human-machine interaction seems to be achievable. Hence, it provides a testbed where it is possible to develop the 'soft skills' needed to sustain such interactions.

Acknowledgement The research reported here has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE)

5 References

Abassi, A.R., Uno, T., Dailey, M., Afzulpurkar, N.V. (2007) Towards knowledge-based affective interaction: situational interpretation of affect. In A. Paiva, R. Prada and R. Picard (eds) *Affective Computing and Intelligent Interaction*, Lisbon, September 2007. Berlin: Springer LNCS pp 452-463.

Aubergé, V., Audibert, N., and Rilliard, A.. (2004) E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. 4th LREC, 179-182, 2004.

Bachorowski, J. A. (1999) Vocal expression and perception of emotion. *Current Directions in Psychol Science* 8(2), 53–57.

Batliner, A., Hacker, C., Steidl, S., Noth, E., and Haas, J. (2003) User States, User Strategies, and System Performance: How to Match the One with the Other. In Proc. ISCA workshop on error handling in spoken dialogue systems, pages 5–10, Chateau d'Oex. ISCA.

Batliner, A., Fischer, K., Huber, R., Spilker, J., & No¨th, E. (2003) How to find trouble in communication. *Speech Communication* 40, 117–143.

Bechara, A., Damasio, A., Damasio, H., & Anderson, S. (1994) Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 7–15.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E.,Sawey M., Schroeder, M., 2000 'FEELTRACE': An instrument for recording perceived emotion in real time.In: Proc. ISCA ITRW on Speech and Emotion, Newcastle, N. Ireland, September 5–7, 2000 pp. 19–24.

Fragopanagos, N & Taylor, J. (2005) Emotion recognition in human-computer interaction. *Neural Networks* 18, 389-405.

Ioannou, S., Raouzaiou, A., Tzouvaras, V., Mailis T., Karpouzis, K., Kollias, S. (2005) Emotion recognition through facial expression analysis based on a neurofuzzy Network. *Neural Networks* 18, 423-435

Kollias, S. et al. (2008) HUMAINE IST 507422 Final report for WP4 (www.emotion-research.net)

Martin, J. C., Devillers, L., Zara, A., Maffiolo, V. and LeChenadec, G. (2006) The EmoTABOU Corpus. Humaine Summer School, Genova, Italy, September 2006

van Reekum, C. M., Johnstone, T., Banse, R., Etter, A., Wehrle, T., & Scherer, K. R. (2004) Psychophysiological responses to appraisal dimensions in a computer game. *Cognition and Emotion* 18, 663–688.

Wang, N., & Marsella, S. (2006). Evg: an emotion evoking game. Proc. 6th International Conference on Intelligent Virtual Agents. Marina del Rey, CA, USA Berlin: Springer LNCS 282-291.

Weizenbaum, J. 1996. ELIZA - A computer program for. the study of natural language communication between man and machine. *Comm. ACM* 9:36-45.

The NIMITEK Corpus of Affected Behavior in Human-Machine Interaction

Milan Gnjatović, Dietmar Rösner
Otto-von-Guericke-University Magdeburg
Department of Knowledge Processing and Language Engineering
P.O. Box 4120, D-39016 Magdeburg, Germany
E-mail: gnjatovic|roesner@iws.cs.uni-magdeburg.de

Abstract

This paper presents the NIMITEK corpus of affected behavior in human-machine interaction. It contains 15 hours of audio and video recordings produced during a refined Wizard-of-Oz (WOZ) experiment designed to induce emotional reactions. Ten native German speakers participated in the experiment. The language used in the experiment was German. During the process of collecting the corpus proper attention was devoted to the issue of its ecological validity. Besides the fact that the refined WOZ simulation gave the opportunity to control development of the dialogue, the problem of role-playing subject was also successfully addressed. The evaluation of the corpus with respect to its emotional content demonstrated a satisfying level of ecological validity. We summarize evaluation results in the following points. The corpus contains recordings of genuine emotions that were overtly signaled. It is not oriented to extreme representations of a few emotions only but comprises also expressions of less intense emotions. Emotional expressions of diverse emotions are extended in modality (voice and facial gesture) and time. In addition, different classes of non-neutral talking style are marked in the obtained data.

1. Introduction

It is a widely accepted fact that research on the role of emotions in human-machine interaction (HMI) is essentially supported by corpora containing samples of emotional expressions. For example, Douglas-Cowie et al. (2004) define the kind of corpus that is needed to support the development of emotion-sensitive interfaces and assess what has been achieved in the field. They emphasize the ecological validity of corpora as one of fundamental requirements – collected samples should be representative of emotions as they occur in everyday life (Douglas-Cowie et al. 2004, p.7). However, they conclude that this requirement is not adequately addressed in existing corpora. Their criticism is leveled against the often used practice of using material produced by actors and disregarding less intense emotions (Douglas-Cowie et al. 2004, p.6-7).

The essence of the problem of assessing the phenomenon of affective behavior as it naturally occurs lies on the methodological level. The question that remains to be open is how to collect such corpora. To illustrate this claim, let us make a simplification of our long-term research aim – we want to develop a spoken natural language dialogue system that should be able to perform two tasks: (1) to determine, based on the recognition of negative user's emotional states, critical phases in interaction and (2) to resolve problems emerged in communication by applying an appropriate dialogue strategy. It should be kept in mind that Brewer emphasizes that we cannot speak of the validity or invalidity of research per se – *validity must be evaluated in light of the purpose for which the research was undertaken in the first place* (2000, p.3). Thus, for the former demand, it is preferable that material contained in the corpus was collected from people experiencing genuine emotions rather than produced by actors. For example, Batliner et al. (2000) show that classification results of a statistical prosodic classifier for emotion recognition from user's spoken input may depend to a high extent on the fact whether it was trained on genuine or acted expressions of emotions. To satisfy this demand for genuine emotions in laboratory settings, it is crucially important (1) to address the problem of role-playing subjects. To satisfy the latter demand, it is useful if the corpus contains samples of dialogues between the user and the system that provide insight in various dialogue strategies that could be applied in order to resolve problems emerged in communication. In other words, researchers should have the possibility (2) to control development of the dialogue between the subjects and the system during the collection of samples.

2. Collecting the NIMITEK corpus

Gnjatović and Rösner (2006) address the methodological desiderata in obtaining a corpus of affected speech in HMI. They propose a refinement of the Wizard-of-Oz (WOZ) technique that meets the two aforementioned requirements to obtain ecologically valid data. The NIMITEK multimodal corpus of affected speech and accompanying facial expressions is collected in the framework of such a refined WOZ simulation. Subjects in the WOZ experiment were asked to undertake a test of both intelligence and communication abilities supported by the spoken natural language dialogue system. In fact they were confronting a set of graphically based tasks specified with the intention to stimulate the verbal interaction between subjects and the system. Tasks were successively displayed on the screen with accompanying descriptions spoken by the system. In order to force subjects to verbally interact with the system, they were only allowed to give spoken instructions to the system. To determine and to formulate acceptable instructions and questions was imputed to be a part of the test as an additional stimulus for subjects to express themselves

verbally. Stimuli used for an emotional response were e.g., intentional misunderstanding of subject's request and performing an incorrect operation, pretending not to understand subject's request and asking for a repetition, confronting subjects to unsolvable tasks, capturing subject's image and displaying it as a part of graphical puzzles, etc. Ten healthy native German speakers (7 female, 3 male) in the age from 18 to 27 (mean 21.7) participated in the experiment. Almost 15 hours of session time were recorded. The language used in the experiment was German.

3. Evaluation of the NIMITEK corpus

We evaluated the NIMITEK corpus with respect to requirements for ecological validity introduced by Douglas-Cowie et al. (2000, p.39-40). Three types of evaluators participated in this process. The first group (three German native speakers) was allowed only to hear audio recordings. These evaluators were influenced by lexical meaning as well. The second group consisted of three non-German speakers: two Serbian native speakers and one Hungarian native speaker (however, the last evaluator was born and living in Serbia, attending schools in Serbian language, etc.). These evaluators did not have knowledge of German language, have never lived in a German speaking environment, and did not have any contact with German language in everyday life. This group was also allowed only to hear audio recordings, but for this group the lexical meaning was missing and thus the prosody became central for evaluating emotions. Finally, one additional German native speaker was allowed to simultaneously hear and see video recordings. Four randomly selected sessions were evaluated in complete duration (approximately five hours), in order that evaluators take the history of interaction into account. The evaluation unit was a dialogue turn or a group of several successive dialogue turns. Only subjects' expressions were evaluated, while wizard's expressions were ignored. The total number of evaluated units was 424.

Evaluators performed this perception test independently from each other. They were given a starting set of "basic" labels {joy, sadness, anger, fear, disgust, neutral}, but they were also allowed to extend this set with additional labels, if necessary, according to their own perception. To each evaluation unit evaluators assigned one or more labels. Recordings evaluated as emotional were further graded with respect to their intensity (three different levels: low, medium, high). Introduced labels are classified in three groups:

- Emotion labels,
- Subject's state labels,
- Talk style labels.

We used majority voting in order to attribute labels to evaluation units. Table 1 shows all the introduced labels and the numbers of cases with majority voting for the first two groups of evaluators (i.e., German native speakers and non-German speakers that were allowed only to hear audio recordings). We consider two kinds of majority voting:

- weak majority – exact two evaluators in a group agreed,
- strong majority – all three evaluators in a group agreed.

A total number of cases with majority voting is the sum of numbers of cases with weak and strong majority voting. Labels with no majority voting (i.e., Fear and Disgust) are also given in the table.

Labels	German speakers majority voting			non-German speakers majority voting		
	total	weak	strong	total	weak	strong
Emotions						
Anger	77	46	31	18	12	6
Nervousness	8	8	-	**224**	131	93
Sadness	8	7	1	1	1	-
Joy	17	14	3	1	1	-
Contentment	12	12	-	4	4	-
Boredom	9	5	4	13	10	3
Fear	-	-	-	-	-	-
Disgust	-	-	-	-	-	-
Neutral	**205**	124	81	54	45	9
Subject's state						
Interested	2	1	1	1	1	-
Surprised	6	4	2	22	16	6
Insecure	26	19	7	71	60	11
Disappointed	17	14	3	12	12	-
Impatient	35	32	3	-	-	-
Confused	30	21	9	-	-	-
Accepting	8	5	3	-	-	-
Pleased	2	2		-	-	-
Stressed	47	43	4	-	-	-
Thinking	10	10	-	-	-	-
Talk style						
Commanding	53	47	6	137	94	43
Off-talk	59	38	21	94	40	54
Pedagogical	15	10	5	73	60	13
Ironic	4	4	-	36	34	2

Table 1: Introduced labels and majority voting.

It should be mentioned that the main aim of this evaluation phase was to demonstrate the *naturalness* of the collected recordings. Thus, selected evaluation units are rather long in duration. Such units were selected to demonstrate that emotional expressions are extended in time. The second evaluation phase was performed for the purpose of defining a user state model that is appropriate for implementation of a prototype spoken dialogue system. We used finer selection of units—the same evaluation material was divided in 2720 evaluation units. For details please see Gnjatović and Rösner (2008b).

4. Discussion

The evaluation process demonstrated a satisfying level of ecological validity of the NIMITEK corpus. It is worth mentioning several points.

(1) Subjects signaled emotions overtly. Confrontation to

the simulated test proved to be a strong motivational factor. The combination of a motivating environment with already mentioned additional stimuli for an emotional response induced subjects to signal their emotions overtly. The evaluation of the NIMITEK corpus shows that subjects signaled both positive and negative emotions. However, induction of negative emotions and emotion-related states was significantly more effective than induction of positive emotions and emotion-related states. The group of *positive* labels contains *joy*, *contentment* and *pleased*, as well as *interested* and *thinking* that can be considered positive with respect to subject's engagement to solve the given task. The other labels, except *neutral*, belong to the group of *negative* labels. According to majority voting results, German speaking evaluators attributed 10.14% of evaluations units with a positive label and 63.92% with a negative label, while non-German speaking evaluators attributed 1.42% of evaluations units with a positive label and 85.14% with a negative label.

(2) Diversity of signaled emotions and their intensities. Another property of the corpus is that it is not oriented to extreme representations of a few emotions only (for example: *anger*, *joy*, *fear*, etc.), but comprises also expressions of less intense, not *full-blown*, emotions (for example: *nervousness*, *pleased*, *insecure*, etc.). A convincing illustration of this fact is that non-German evaluators attributed 52.83% of evaluation units with *nervousness*, and only 4.25% with *anger*. As mentioned above, units attributed by an emotional label are further graded with respect to the intensity of signaled emotion. Table 2 shows the numbers of assigned emotional labels classified by intensity. In this table, we do not resort to majority voting, but give the absolute numbers of assigned labels.

Emotion	***Low***	***Medium***	***High***
Anger	248	144	30
nervousness	466	270	33
Sadness	26	27	1
Joy	66	41	3
contentment	130	23	3
Boredom	142	27	1
Fear	8	11	-
Disgust	6	-	-

Table 2: Numbers of assigned emotional labels classified by intensity of expressed emotion.

(3) Emotional expressions are extended in modality. Our experiment deals with the expressions of emotions in two modalities at a time: vocal and facial expressions. Although vocal expressions are prioritized in our research, such settings give an opportunity to observe the correlation between these two modalities.

(4) Emotional expressions are extended in time. At the level of emotions, this is important because different development phases of emotional expressions could be observed. In this sense, the prosodic (suprasegmental) realization of affected speech is especially indicative. Moreover, at the language level, collection of larger units is also valuable, because they function more directly in the realization of higher-level patterns (Halliday, 1994, p.19).

(5) Additional shared non-linguistic context. In our experimental settings, the desktop of the subject's PC is also recorded. It represented an additional non-linguistic context shared between subjects and the simulated system. Subjects considered it to be a reliable source of information. In such cases when wizard's actions were in a collision with the actual state on the desktop, subject's often tried to refer first to the desktop. The inspection of all 6772 commands spontaneously produced by subjects shows that non-linguistic context influenced the language of subjects to a high extent with respect to frequency of "irregular" (e.g., elliptical or minor, etc.) utterances.

(6) Different classes of non-neutral talking style. Four classes of non-neutral talking style are marked in the obtained data: *commanding*, *off-talk*, *pedagogical*, *ironic*. Although these classes differ in the level of interactivity, they all carry information about speaker state and intention.

5. Conclusion

This paper presents the NIMITEK corpus of affected behavior in human-machine interaction. It contains 15 hours of audio and video recordings produced during a refined Wizard-of-Oz (WOZ) experiment designed to induce emotional reactions. Ten native German speakers participated in the experiment. The language used in the experiment was German.

During the process of collecting the corpus proper attention was devoted to the issue of its ecological validity. Besides the fact that the refined WOZ simulation gave the opportunity to control development of the dialogue, the problem of role-playing subject was also successfully addressed. The evaluation of the corpus with respect to its emotional content demonstrated a satisfying level of ecological validity. We summarize evaluation results in the following points. The corpus contains recordings of genuine emotions that were overtly signaled. It is not oriented to extreme representations of a few emotions only but comprises also expressions of less intense emotions. Emotional expressions of diverse emotions are extended in modality (voice and facial gesture) and time. In addition, different classes of non-neutral talking style are marked in the obtained data.

Finally, we briefly mention some lines of research that were supported by the NIMITEK corpus.

(1) Modeling attentional information. Inspection of the NIMITEK corpus showed that subjects often produced "irregular" (e.g., elliptical or minor, etc.) utterances. Thus, there was a need to develop structures and algorithms that support system's decision making processes when it is confronted with such user inputs. Gnjatović and Rösner (2007a) introduce the concept of the focus tree in order to model attentional information on the level of a user's command and the rules for transition of the focus of attention for different types of user's commands.

(2) Introducing dialogue strategy for supporting users. Resorting to the NIMITEK corpus, Gnjatović and Rösner (2007b, 2008a) introduce an adaptive dialogue strategy for supporting users while they solve a graphical task. Aimed to address negative user states in human-machine interaction, this dialogue strategy supports users when they have problems related to the task itself or to the interface language. The central idea is that the dialogue strategy is dynamically refined according to the current state of the interaction.

(3) Implementing the NIMITEK prototype system. Two above mentioned directions of our research were integrated in the conceptual design and implementation of the NIMITEK prototype system. It is a spoken dialogue system for supporting users while they solve problems in a graphics system. The role of the NIMITEK corpus in the development of the prototype system is discussed by Gnjatović and Rösner (2008b).

(4) Ongoing research. In an ongoing experiment, the audio recordings from the NIMITEK corpus are used as stimuli in a functional magnetic resonance imaging (fMRI) study of prosody processing. This study is expected to provide a better insight in how users percept prosodically marked spoken output of the system.

Please note that the NIMITEK corpus is available from the authors for research purposes upon request.

6. Acknowledgements

The presented study is performed as part of the NIMITEK project (http://wdok.cs.uni-magdeburg.de/nimitek), within the framework of the Excellence Program "Neurowissenschaften" of the federal state of Sachsen-Anhalt, Germany (FKZ: XN3621A/1005M). The responsibility for the content of this paper lies with the authors.

7. References

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E. (2000) Desperately Seeking Emotions: Actors, Wizards, and Human beings. *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, p.195-200.

Brewer, M. (2000) Research Design and Issues of Validity. Reis, H. and Judd, C. (eds.) *Handbook of Research Methods in Social and Personality Psychology*. Cambridge:Cambridge University Press, p.3-16.

Douglas-Cowie, E., Cowie, R., Schröder, M. (2000) A new emotion database: Considerations, sources and scope. *Proceedings of the ISCA Workshop on Speech and Emotion*, p.39-44.

Douglas-Cowie, E., WP5 members (2004) Preliminary plans for exemplars: Databases. Public Report, HUMAINE Project, http://emotion-research.net/deliverables/D5c.pdf.

Gnjatović, M., Rösner, D. (2007a) An approach to processing of user's commands in human-machine interaction. *Proceedings of the 3rd Language and Technology Conference (LT&C'07)*. Adam Mickiewicz University, Poznan, Poland, pages 152--156.

Gnjatović, M., Rösner, D. (2007b) A Dialogue Strategy for Supporting the User in Spoken Human-Machine Interaction. *Proceedings of the XII International Conference "Speech and Computer" (SPECOM 2007)*. Moscow State Linguistic University, Moscow, Russia, pages 708--713.

Gnjatović, M., Rösner, D. (2008a) Emotion Adaptive Dialogue Management in Human-Machine Interaction. *Proceedings of the 19th European Meetings on Cybernetics and Systems Research (EMCSR 2008)*. Austrian Society for Cybernetic Studies, Vienna, Austria, pages 567--572.

Gnjatović, M., Rösner, D. (2006) Gathering Corpora of Affected Speech in Human-Machine Interaction: Refinement of the Wizard-of-Oz Technique. *Proceedings of the International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS 2006)*, Academia Sinica, Taipei, Taiwan, p.55-66.

Gnjatović, M., Rösner, D. (2008b): On the Role of the NIMITEK Corpus in Developing an Emotion Adaptive Spoken Dialogue System. *Proceedings of the 6th edition of the Language Resources and Evaluation Conference (LREC 2008)*. European Language Resources Association (ELRA). Marrakech, Morocco. To appear in May 2008.

Halliday, M.A.K. (1994) An introduction to functional grammar, Second Edition, London, Edward Arnold.

An Interface to Simplify Annotation of Emotional Behaviour

Shazia Afzal, Peter Robinson
Computer Laboratory, University of Cambridge
15 JJ Thompson Avenue, Cambridge, CB3 ODF, UK
E-mail: Shazia.Afzal@cl.cam.ac.uk, Peter.Robinson@cl.cam.ac.uk

Abstract

Research in affective computing is increasingly moving towards naturalistic data. Capturing and annotating such complex data is a massively challenging task. This paper describes a simple and efficient annotation scheme that promotes context-sensitive data labelling via an easy to use interface in an attempt to reduce reliance on expert or trained coders. Additionally, the same labelling interface can be used to obtain self-report of emotional behaviour from subjects. This annotation method has been designed to allow faster labelling of data with a minimal learning curve as part of our research in studying non-verbal expressivity of affect in computer based learning environments and is currently being evaluated. Design decisions are based on feedback from usage of initial prototype as well as relevance to our domain of interest. We anticipate that this can enable faster preparation of representative data in an effective manner for use in automatic analysis studies.

1. Introduction

As affect (or emotion) research gradually integrates with HCI studies and matures in application from mere prevention of usability problems to promoting richer user experiences, the need to capture 'pervasive emotion' (Cowie et al., 2005) and also its context of occurrence is becoming an increasing concern. Our research involves modelling affective aspects of learner experience in computer assisted learning environments. As such we are interested in studying how non-verbal behaviour from multiple-cues like facial expressions, eye-gaze and head posture can be used to infer a learner's affective state during interaction and learning with a computer tutor. The ultimate objective is to abstract this behaviour in terms of features that can enable automatic prediction and reliable computational modelling of different affect states. The need for representative data is therefore essential in order to carry out realistic analysis, to develop appropriate techniques and eventually perform validation of inferences.

Capturing naturalistic data - as it occurs and in all its complexity, is however a massively challenging task. Existing databases are often oriented to prototypical representations of a few emotional expressions, being mostly posed or recorded in scripted situations. Such extreme expressions of affect occur rarely, if at all, in HCI contexts. The applicability of such data therefore becomes severely limited because of observed deviation from real-life situations (Batliner et al., 2003) and for our purpose, their relevance to a learning situation like one-on-one interaction with a computer tutor. For developing systems that generalise to real world applications there is now an increasing shift from easier to obtain posed data to more realistic naturally occurring data in the target scenarios. Dealing with the complexity and ambiguity associated with natural data is however a significant problem.

Automatic prediction using machine learning relies on extensive training data which in this case implies preparation of labelled representative data. This also serves as a ground-truth for validation of developed techniques and is therefore a crucial necessity. Non-verbal behaviour is rich, ambiguous and hard to validate making labelling of data a tedious, expensive and time-consuming exercise. In addition, lack of a consistent model of affect makes the abstraction of observed behaviour into appropriate labelling constructs very arbitrary. To achieve a compromise between descriptive detail and economy of annotation effort as in Kipp et al. (2007), this paper describes an annotation scheme tailored to our research but also applicable to similar areas. It is designed to map spontaneous interpretation of recorded behaviour onto different affect states and is currently being evaluated.

In Section 2 we describe the annotation method along some parameters that we deem to be important while considering an annotation scheme. Section 3 discusses some limitations and possible improvements to enhance the procedure while Section 4 concludes the paper by summarising the main idea.

2. Annotation Method

The annotation method that we describe evolved from various domain relevant decisions related to the choice of labelling constructs and modality, anticipated technical constraints in target scenario, relation to context and ease of interpretation. It is inspired by socially-based coding schemes; that is, observational systems that examine behaviour or messages that have more to do with social categories of interaction like smiling rather than with physiological elements of behaviour like amplitude (Manusov, 2005). Precisely, Bakeman & Gotham (1997) define a socially based scheme as one "that deal with behaviour whose very classification depends far more on the mind of the investigator (and others) than on the mechanisms of the body".

The scheme is designed to allow a split-screen viewing of a subjects' recorded behaviour along with the time synchronised interaction record obtained via screen capture. It is implemented in the form of an easy to use annotation interface that combines viewing, navigation and labelling of recorded data. Figure 1 below shows a snapshot of a labelling session using the interface.

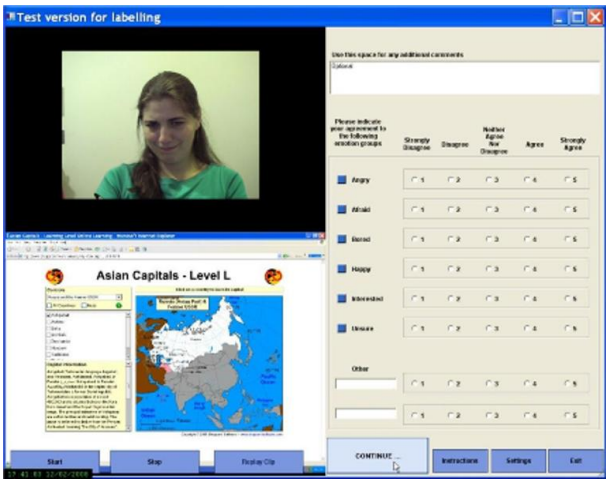


Figure 1: Snapshot of the annotation interface

The annotation scheme and different features are described here along the following parameters.

Labelling Constructs: Decisions related to the representation of affect permeate every subsequent step in the automatic analysis of non-verbal behaviour. Annotation schemes commonly employ either categorical, dimensional or appraisal based labelling approaches. In addition, free-response labelling may also be used for subjective descriptions. For a description and relative merits of each method the interested reader is referred to (Cowie et al., 2005; Douglas-Cowie et al., 2004).

We are using a variant of categorical labelling where coders are asked to rate their agreement on each of the pre-selected categories based on a Likert scale ranging from Strongly Agree to Strongly Disagree. The categories reflect the macro-classes of a taxonomy of complex mental states selected for their relevance to the domain of study and therefore include affect states that are considered pertinent in learning situations (Afzal & Robinson, 2007).

Getting agreement ratings on all affect descriptors on a single data segment allows a greater degree of freedom in inference tasks. To reduce the bias of forced choice on selected affect labels - an often listed drawback in categorical methods, the scheme allows the coder to define his/her own category or label if the perceived state is not represented by the categories. Additionally, there is provision for a free-form description should the coder wish to include comments or other observations not captured via categorical listing. The flexibility in labelling is provided consciously in order to characterise mixed emotions that are known to occur frequently in realistic settings.

Level of measurement: Observational assessment can be done along two different frames of reference – at a macro level to capture the social meaning of behaviour or at a micro level to analyse specific cues or displays in behaviour (Manusov, 2005). Our purpose is of capturing the affective component in behaviour - which is influenced by social meaning, rather than coding of individual displays like smiles, head gestures, eye-gaze, etc.

Context Information: Expressivity and context interact in complex ways as behaviour is always interpreted within a certain context. To emphasise the significance of context in the perception of meaning Russel (1997) cites the example of an experiment where three silent film strips each ending with the same footage of a deliberately deadpan face of an actor were created. In each strip the face was preceded by a different picture - a bowl of soup, a dead woman in a coffin and a young girl playing with a teddy bear. The result was an illusion – audiences saw emotions expressed in the actors' deliberately posed expressionless face (Russel, 1997). Such varying interpretations based on varying context information indicates the danger of forming judgements in isolation from the context of occurrence. Ignoring context can thus dangerously introduce a relative bias in inferences made solely from non-verbal behaviour (Russell J.A., Fernandez-Dols, 1997; Jinni et al., 2005).

In order to represent context information we explicitly try to recreate the interaction by making available both the activity and the user views so that the coder does not need to spend additional time in 'creating' a context. This coupling of information recreates the evolution of behaviour on task and primes the coder into making context-sensitive judgement. The idea is contextualisation of meaning by combined assessment.

Coding Unit: The coding unit refers to the decisions regarding when to code within the interaction and the length of time the observation should last (Manusov, 2005). It has two broad variants - event based and interval based. The choice of the coding unit depends upon the research view and the level of accuracy required, complexity of the coding scheme and the frequency of behaviour occurrence (Bakeman & Gotham, 1997). Our method implements the interval based coding unit through fixed time slots. Also known as systematic observation, this has the advantage of allowing behaviour to be observed consistently throughout the interaction allowing a more accurate reflection of how it is represented in the data (Manusov, 2005).

The initial prototype of the tool implementing the annotation method operated in two modes: manual and timed. In manual mode, the choice of determining when to label was left to the coder while in the timed mode the coder was prompted for noting the annotation after preset durations like every 5 seconds, 10 seconds or 20 seconds. It was observed that non-expert coders preferred the timed

mode over the manual one as being much easier and convenient. Manual operation requires controlled navigation while maintaining the reference to context. Coders felt that it distracted them from observing the sequential behaviour. By not having to define segments of emotional episodes they could focus solely on the observation and hence labelling of behaviour. As such the manual mode was disabled in the current implementation of the annotation scheme and only the fixed interval coding unit was retained.

Dynamic Interpretation: Instead of pre-segmenting video clips for labelling, the method forces labelling in temporal sequence. In this way it retains the natural evolution of the behaviour and preserves the dynamics of expression and interaction.

Level of Expertise & Ease of Use: Selection of coders or raters is important for the labelling process as they should be able to discern meaning from behaviour and make judgements. Reliance on expert or trained coders makes the labelling task very time-consuming and expensive. Since the effectiveness of coders depends hugely on the nature and complexity of the coding system applied (Manusov, 2005), the design of the interface and coding scheme was simplified in an attempt to include non-experts as coders. Annotation tools like FEELTRACE (Cowie et al., 2000) and ANVIL (Kipp, 2001) require considerable training before use and restrict access to expert coders owing to the associated learning curve. Our proposed annotation method can on the other hand be used by diverse people without prior experience in labelling. To ensure quality of observation however, the coders can be pre-tested on their nonverbal decoding ability. Initial evaluations show that users are able to perform labelling smoothly soon after being familiarised with the interface and labelling procedure.

Self-Reporting: Inter-coder agreement scales like Cohen's Kappa are used for validation of annotation but are highly sensitive to the affect decoding skills and gender of individual coders (Abrilian, 2005). Obtaining self-report from subjects is an effective strategy of cross validation and interpretation of behaviour. Usage of standard self-report instruments like SAM (Lang, 1980) and EmoCards (Desmet et al., 2001) depends on specific research setups and factors like type of data sought, resources available, situation and users (Isomursu et al., 2007). Our method allows ease in comparison since the same interface used for labelling by external coders can also be used to obtain self-report. Verbal feedback from subjects using this method for self-reporting verified the utility of providing context knowledge and also the ease in usage. Of interest here is that even while self-reporting affect judgements, subjects preferred to work in the fixed interval timed mode rather than event based mode.

Optimisation of annotation effort: The method economises annotation effort by eliminating the need to iterate over data for hierarchical labelling as proposed Abrilian et al. (2005). The structure of the labelling format implicitly incorporates the elements of multi-step or hierarchical annotation as recommended.

Output Format: Each labelling session produces annotations in exportable *csv* or *xml* files. This allows seamless integration with data analysis tools and hence faster interpretations.

3. Limitations & Possible Extensions

Use of pre-selected categorical labels is an unavoidable limitation and has been done to cater to our domain of study. Also, dimensional constructs like valence have a relative meaning. Confusion, for instance, is considered a negatively valenced emotion and but has been found to have a positive effect on learning (Craig et al., 2004). So if a coder has to label the valence of a specific behaviour it will be difficult to establish whether the valence represents the objective view per se or is to be understood in relation to the current task.

Another drawback of our approach is that it will fail to account for emotional transitions occurring at the periphery of the fixed time intervals for observation. Depending on the frequency of such occurrences this can be easily overcome by repeating the annotation on a different time-scale. Interpreting results on the same source labelled on different time scales is trivial as the larger time grain can always be defined in terms of the smaller time segments and thus easily compared.

Further extensions to improve the annotation mechanism involve inclusion of context attributes like theme, degree of implication, target of emotion, communicative goal and the cause of emotion (Abrilian, 2005). Additions of more labelling attributes will however increase the complexity and difficulty of the labelling process. Online availability of the annotation tool to facilitate access and coordinate the labelling process is also proposed.

4. Summary & Conclusions

Labelling of data has a dual purpose. For computational analysis it serves as a ground-truth for evaluation and comparison of performance. More importantly, it serves as a key knowledge source to develop an understanding of affective behaviour that may occur in a learning situation and how it is perceived by humans. It is non-trivial in terms of the complexity associated with deciding the correct representation and descriptors of emotional behaviour as well as in the overall effort required for the task. Further, sensitivity of emotions to the form of measurement makes it more challenging to arrive at an optimal annotation format. Since quality of annotated data determines the efficiency of automatic prediction techniques, the choice of an annotation methodology is an important determinant of the true usefulness of collected video data.

This paper describes a simple and yet effective annotation method that can be easily administered to allow faster labelling of naturalistic data. The motivation to develop a simplified interface for annotation was to include non-experts in the coding process and utilise their general skills of decoding nonverbal behaviour. The annotation scheme is designed as part of our study of non-verbal behaviour in learning environments and is being evaluated.

5. Acknowledgements

This research is supported by the Gates Cambridge Trust and the Overseas Research Studentship of the University of Cambridge.

6. References

Abrilian, S., Devillers, L., Buisine, S., Martin, J-C (2005). "EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces", *HCI International*.

Afzal S., Robinson P. (2007). A Study of Affect in Intelligent Tutoring, In *Proceedings of Workshop on Modelling and Scaffolding Affective Experiences to Impact Learning, International Conference on Artificial Intelligence in Education*, Los Angeles.

Bakeman R., Gothman J.M. (1997). Observing interaction: An introduction to sequential analysis, Cambridge University Press, UK.

Batliner A., Fischer K., Huber R., Spilker J., Noth E. (2003). How to Find Trouble in Communication, Speech Communication, 40 (1-2), pp. 117-143

Boener K., DePaula R., Sourish P., Sengers P. (2007). How emotion is made and measured, *Int. J. Human-Computer studies*, 65, pp. 275-291

Cowie, R., Douglas-Cowie E. & Cox C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks, *Neural Networks*, 18, 371-388

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time. In *ISCA Workshop on Speech & Emotion*, (pp. 19-24). Northern Ireland.

Craig, S.D., Graesser A.C., Sullins, J. & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29, pp. 241-250.

Desmet P.M.A., Overbeeke P.J., Tax S.J.E.T. (2001). Designing products with added emotional value; development and application of an approach for research through design, *The Design Journal* **4** (1), pp. 32–47.

Devillers L., Vidrascu L., Lamel L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18, pp. 407-422.

Douglas-Cowie, E. et al. (2004). Deliverable D5c-Preliminary plans for exemplars: databases. *Project Deliverable of Humaine Network of Excellence*.

Isomursu M., Tahti M., Vainamo S., & Kuutti K. (2007). Experimental evaluation of five methods for collecting emotions in field settings with mobile application, *Int. J. Human-Computer Studies*, 65, pp. 404-418.

Jinni, A. Harrigan, J. A., Rosenthal, R., and Scherer, K. R. (2005). *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of Eurospeech'2001*.

Kipp M., Neff M., Albrecht I. (2007). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. *Journal on Language Resources and Evaluation-Special Issue on Multimodal Corpora*.

Lang P.J. (1980). Behavioral treatment and bio-behavioral assessment: computer applications. In: J.B. Sidowski, J.H. Johnson and T.A. Williams, Editors, *Technology in Mental Health Care Delivery Systems*, Albex, Norwood, NJ (1980), pp. 119–139.

Manusov V.L. (2005). *Sourcebook of Nonverbal Measures: Going Beyond Words.* Lawrence Erlbaum Associates.

Russell J.A., Fernandez-Dols J.M. (1997). *The psychology of facial expression*. Cambridge, MA: Cambridge UP.

Anger detection performances based on prosodic and acoustic cues in several corpora

Laurence Vidrascu and Laurence Devillers

LIMSI-CNRS
BP 133, 91403 Orsay cedex
{devil, vidrascu}@limsi.fr

Abstract

Anger detection is a relevant technology for improving call center applications, but requires an emotion detection system with a high level of performances for different tasks. This paper deals with a study of anger detection system performances based on prosodic and acoustic cues in several corpora with felt and portrayed emotions. The ground question of this paper is that of the portability of emotion detection systems. What are the performances of models trained on specific real-life data and tested on other data such as portrayed data? We show that our emotion detection system, which was built on a real-life call center corpus, can well detect Anger from a different call center corpus and can also detect Anger from a portrayed data corpus.

1. Introduction

Emotion detection and especially anger detection could help improve call center applications. Yet, when anger is studied in many classification tasks, it is not clear that what different studies call "anger" refers to the same state and thus that a detection system trained in a specific context might perform well on other data.

Anger is one of the primary emotions defined by Darwin [1] and is considered in a consensual way by the scientific community as a primary emotion. For Lazarus [2], the core relational theme of anger is "a demeaning offence against me and mine". For Shaver [3], it is "something that interferes with the person's attainment of certain goals. The angry person makes the perception that the harm is illegitimate [..] (the) situation is contrary to what ought to be". The first cognitivist theoretician of emotions is in fact Aristotle who, in a surprisingly modern way, defines anger in Rhetoric [4], as "an impulse, accompanied by pain, to a conspicuous revenge for a conspicuous slight directed without justification towards what concerns oneself or towards what concerns one's friends."

Although specific expressive patterns are theoretically a feature of primary emotions, Ekman suggests that the category 'anger' can be conceived as family, with members sharing common features (such as expressive universals and invariants) but also displaying variations, depending on individuals and/or specific circumstances. Brenner [5] further points out, several manifestations of the same "primary emotion" may result in fact from very different appraisals which lead to variable expressions.

Scherer [6] distinguishes two types of anger: cold and hot anger and cites a study of Frick [7] on two types of anger, one linked to frustration, the other to aggression, which have quite different acoustic manifestations and are perceptively differentiable. In [8], Scherer assesses human performances for the recognition of six emotions (Anger, Fear, Joy, Sadness, Disgust and Neutral state) in voice and face by comparing several studies accomplished with actors. The recognition rates for voice are between 55 % and 65 % with big variations according to the emotions studied, anger and sadness being best recognized with scores often between 70 and 80 %.

In real-life corpora, the term 'anger' can be used to designate a primary emotion, a mood, an attitude or a mixture of these different affective states and it is often mixed with other emotions [9]. In this paper, the widely used term 'emotion' is used without distinction from the more generic term 'affective state', which may be viewed as more adequate from the psychological theory point of view.

The present study investigates the generic power of emotion detection systems in the case of anger. Our purpose is to study the portability of detection system built on a specific corpus recorded in a French call-center and tested first on data recorded in a very similar task and then on data recorded with actors. Corpora of acted portrayals are usually not representative of the range of emotional expressions likely to occur in any specific real-life context. They tend to include portrayals of very intense emotional reactions which are considered to occur infrequently in daily interactions between humans or between humans and machines. Acted portrayals have therefore been challenged as unsuited for applied research purposes. We are aware of the difficulty of the task and of the large number of variables that need to be taken into account for allowing comparisons. Quality of the audio signals, definitions of emotions, annotations of emotions, etc, all vary across corpora and seriously restrict comparability across corpora. But we think that it is important to try to assess the benefits and the drawbacks of different approaches to data collection in order to improve the quality of the computational models and to reduce the cost of the data collection. Our first results must be interpreted with caution, as explained in the result and conclusion sections. In the following sections, we describe the corpora (section 2), the data processing (section 3) and the comparison between the different corpora (section 4). Conclusions and further research are discussed in section 5.

2. Corpora

The study reported in this paper makes use of two corpora of naturally-occurring dialogs recorded in real-life call centers and a sub-set of the GEMEP [10] corpus which is a collection of portrayed emotions. The call center corpora are hand-transcribed and include additional markings for microphone noise and human produce non-speech sounds (speech, laugh, tears, clearing throat, etc.). The use of these data carefully respected ethical conventions and agreements ensuring the anonymity of the callers, the privacy of personal information and the non-diffusion of the corpus and annotations.

2.1. CEMO: French Medical call center corpus

The CEMO corpus contains real agent-caller recordings obtained from a convention between a medical emergency call center and the LIMSI-CNRS. The transcribed corpus contains about 20 hours of data. The service center can be reached 24 hours a day, 7 days a week. Its aim is to offer medical advice. An agent follows a precise, predefined strategy during the interaction to efficiently acquire important information. His role is to determine the call topic and to obtain sufficient details about this situation so as to be able to evaluate the call emergency and to take a decision. This study is based on a 20-hour subset comprised of 688 agent-client dialogs (7 different agents, 784 clients). About 10% of speech data is not transcribed since there is heavily overlapping speech. The description of the emotion annotation strategies are given in [9, 11].

2.2. BOURSE: French Financial corpus

The BOURSE corpus contains real agent-client recordings obtained from a Web-based Stock Exchange Customer Service Center within the context of the Amities project [12]. The service center can be reached via an Internet connection or by directly calling an agent. While many of the calls involve problems in using the Web to carry out transactions (general information, complicated requests, transactions, confirmations, connection failures), some of the callers simply seem to prefer interacting with a human agent. 100 agent-client dialogs (4 different agents) in French were orthographically transcribed and annotated. The dialogs cover a range of investment-related topics such as information requests (services, commission fees, stock quotations), orders (buy, sell, status), account management (open, close, transfer, credit, debit) and Web questions and problems. There are about 6200 speaker turns in this corpus. Our studies [13] are based only on the 5000 speaker turns after the exclusion of overlaps, which are known to be frequent phenomena in spontaneous speech. The description of the emotion annotation strategies are given in [13].

2.3 GEMEP: French Portrayed emotion corpus

GEMEP is a multimodal corpus of acted emotional utterances, recorded by the Geneva Emotion Research Group[1]. Ten professional French-speaking actors – five male and five female – portrayed 15 affective states under the direction of a professional film director. More than one type of *anger*, *fear* and *sadness* (as well as a broad range of *positive* emotions) were included in order to increase the variability of the expressions portrayed and include states that might occur in daily interactions (see Table 1 for the definitions of *hot* and *cold anger*).

Table 1. *Emotion definition for Anger*

Anger (Hot)	Violent dissatisfaction caused by another person's stupid or malicious actions
Irritation (Cold)	Reaction experienced while attempting to remain cold-blooded when confronted with a very annoying event or person

The actors were provided with definitions and scenarios for each affective state and were requested to improvise interactions with the director. For each affective state, they were requested to produce two pseudo-linguistic sentences, a sustained vowel ('aaa'). All actors produced several repetitions of the standard verbal content (the two sentences and sustained vowel), as well as improvised sentences in French. The actors were further requested to regulate the expressions, showing less and more intense emotions, as well as masked (partly suppressed) expressions. A microphone was positioned over the left ear of the actor, providing a separate speech recording with a constant distance to the actor's mouth. Actors did not portray "neutral" ("non-emotional") states, but portrayed a relatively mild affective state labelled 'interest'.

3. Features extraction and classification

Two experiments were made with models trained on the CEMO corpus, one on the corpus BOURSE of stock exchange transaction (call center, telephone data with situation and different behaviors) and the other one on acted data, We used Praat [14] for features extraction and Weka [15] for training computational models.

3.1. Call center data

Stock exchange data (BOURSE) are comparable with CEMO data in so far as they both come from telephone interactions. Experiments were accomplished with SVMs and 116 attributes extracted by segments (F0, formants, energy, micro-prosody, emotional markers). Details about the features are given in [16].

3.2. GEMEP data

For the comparison between CEMO and GEMEP data, we only tested the pseudo-linguistic and improvised sentences in modes normal and intense. GEMEP portrayals are quite different from call center data: the audio sampling is 44kHz, the portrayed states are defined by the actors expressive intentions (whereas the call-center data is labelled by expert listeners). Perceptive ratings of GEMEP data are yet unpublished and were not available at the time of our analyses. We have tried to reduce the technical differences. The GEMEP audio signals have been transformed to be more comparable with the telephone recorded data by:

- Under sampling to pass of 44kHz in 8Hz
- Elimination of low frequencies with a filter crosses band (band phones 300Hz-3.4kHz)
- Addition of a phone background noise (acquired in party of a file CEMO).

[1] For more details on this corpus, please see [10]. For the work in progress contact T. Bänziger (Tanja.Banziger@hig.se) or K. Scherer (Klaus.Scherer@pse.unige.ch)

But this transformation did not have an impact on results.

4. Results

All experiments were accomplished with SVMs and with only acoustic and prosodic attributes extracted by segments (F0, formants, energy, micro-prosody). For the BOURSE corpus we tested both detection systems trained on the CEMO corpus and tested on the BOURSE corpus and vice versa. The corpus CEMO was annotated with about 20 affective states that were grouped in 7 coarse classes (Fear, Anger, Sadness, Relief, Empathy, Positive, Neutral state…). For the comparison between the GEMEP and the CEMO data, we searched for overlapping emotion classes in both corpora and finally settled for a 4 classes detection task (Fear/Anger/Sadness/Relief). Training on the GEMEP corpus and testing on CEMO didn't yield any conclusive results; therefore we chose to only describe experiments that were done using CEMO as train and GEMEP as test.

4.1. Results on Call center data (BOURSE)

A test was first performed for the Neutral/Anger classification, with 3 sets, agents from the CEMO corpus, callers from the CEMO corpus and callers from the BOURSE corpus.

Figure 1. Percentage of recognition on three different test sets of three classifiers trained respectively on CEMO-agent (500 segments per class), CEMO-caller(400 segments per class) and Bourse-Caller (abscissa).

All three sets were divided into a training set and a test set with distinct speakers. A classifier was trained on each of the three sets and then tested on each test set. About 100 acoustical parameters (F0, energy, formants, microprosody) were extracted and the best 20 or 30 were selected for each task (see [9] for a description of the parameters and feature selection).

The results are given in Fig. 1. There is between 75%-80 % of good detection when the train and test are from the same set (80% in CEMO and 75% in BOURSE (with less intense emotion)). As we had seen [10], anger manifestations are different for callers and agents. Thus CEMO callers are better classified (~73%) by the BOURSE Anger/Neutral detection system than by the CEMO agents detection system. The BOURSE data is not so well detected (~66%) by the CEMO Anger/Neutral detection system, but still results are better with the Caller system than the Agent system and all results are above chance.

4.2. Results on GEMEP

As pointed out in other studies [17], performances vary significantly for different actors. We looked at performances by actor with the classifier trained on CEMO data, and withdrew 3 actors (out of ten) who didn't appear to perform as well as the others. Finally, we tested the GEMEP data with a 4-emotion detection system built on CEMO data. The detection system with the four emotions "Anger, Fear, Sadness and Relief" had been used in other studies [12]. On the test set of the CEMO corpus, the results for a 4-emotions detection test (Anger, Fear, Sadness and Relief - caller voices for training and testing) are overall 51% accurate detection using only acoustic and prosodic features (affect bursts are for instance not included in this features set). For GEMEP, we tested the emotions 'hot anger' and 'irritation' that we expected to correspond to the CEMO Anger, sadness and despair for Sadness, worry and fear for Fear and relief for Relief.

	Fear	**Anger**	**Sadness**	**Relief**
wor (394)	**4**	11	71	13
fea (135)	**10**	37	38	15
irr (216)	9	**7**	75	8
ang (127)	13	**61**	9	17
sad (201)	6	2	**81**	10
des (157)	11	36	**45**	8
rel (250)	20	6	70	**4**

Figure 2. Confusion matrix (in percent) for GEMEP emotion categories with a 4-emotion detection system trained on call center data. Irr: irritation, ang: hot anger, sad: sadness, des: despair, wor: worry/anxiety, fea: panic fear, re: relief. The number of emotional segments is given in parenthesis. The number into parenthesis gives the number of samples per category)

The results for the GEMEP test with 7 actors (portrayals with normal and high level of intensity were selected) are given on Fig.2. Eighty percent of the 'sad' portrayals are accurately classified, but all other low-aroused emotion categories ('worry', 'relief' and 'irritation') are also predominantly classified as Sadness (according to the CEMO trained model). Anger is detected with 61 % accurate recognition; irritation is almost never detected as Anger. The two remaining high-aroused emotion categories in GEMEP ('despair' and 'panic fear') are also predominantly classified as 'anger' or 'sadness'. GEMEP portrayals overall are seldom classified in the CEMO models for 'fear' and 'relief'. Besides, anger seems to be better detected in sentences with pseudo-linguistic standard content probably because everything must be encoded in the prosody (about 67 % of recognition of anger for the 2 first sentences, against 51 % for improvisation).

5. Conclusions

This paper raises the meaningful question of how close to the intended context, the training material need to be. From an application point of view, it is very interesting

to use acted data for training; but is that realistic? Our first analysis tends to show that a 4-emotion (Anger, Fear, Sadness and Relief) detection system built on real life call center data does not allow recognizing the same emotions in GEMEP acted data. Only Hot Anger was detected, most of the other classes being recognized as Sadness. Anger is also well detected for another call center task. The conclusion is that for some emotions such as Hot Anger, some archetypal expressions may exist whatever the context. For other emotions such as Fear, or Irritation, the expressions seem to be more dependent on different definitions within different corpora, or on appraisal events and/or on contextual information. Banziger et. Al [10] have argued that: (a) portrayals produced by appropriately instructed actors are analogue to expressions that do occur in selecte real-life contexts; (b) acted portrayals – as opposed to induced or real-life sampled emotional expressions – display the most expressive variability. It is difficult to draw conclusions about point (a) after our comparison between the CEMO and the GEMEP data, especially taking into consideration the facts that CEMO is task and context dependant and that affective states definition depend on the different actors for GEMEP. Yet, we argue after our first analysis that without any task in mind, the portrayed emotion might not be useful for building real-life system except for some intense emotions such as Hot Anger. However, our first results must be interpreted with a lot of precautions. A remaining question is the fact that the difference between both corpora (GEMEP and call center) could be also due to the multimodality. Other studies such as human perceptive test comparing portrayed expressions to expressions sampled in various real-life contexts are needed in order to conclude on assertion (a) and investigate assertion (b).

Acknowledgements

This work has been conducted in the framework of cooperation between the NCCR in Affective Sciences (UNIGE) and the LIMSI-CNRS in the NoE HUMAINE. The authors would like to especially thank Tanja Bänziger and Klaus Scherer for sharing data and inputs to this manuscript.

6. References

[1] C. Darwin, The expression of the emotions in man and animals, *John Murray*, London, 1872.

[2] Lazarus, R. S. (1998). *Fifty years of the research and theory of R.S. Lazarus*.

[3] Shaver, P., J. Schwartz, D. Kirson and C. O'Connor (2001). Emotion knowledge: Further Exploration of a Prototype Approach. *Emotions in social psychology*. W. Parrott. Philadelphia, Psychology Press: p: 26-56.

[4] Aristotle (Rethorique II 1378a).

[5] Brenner, C. (1980). A psychoanalytic theory of affects. Emotion theory, research and experience vol1. R. Plutchik and H. Kellerman. New York, Ac. Press: p: 341-348.

[6] Banse, R. and Scherer. K. R. (1996). Acoustic profiles in vocal. emotion expression. J. Pers. Soc. Psychol. 70 (3), 614-636.

[7] Frick, R. W. (1986). The prosodic expression of anger: Differentiating threat and frustration. *Aggressive Behavior*. **12**: p. 121–128.

[8] Scherer, K. R. (2003). Vocal communication of emotions : A review of research paradigm. *Speech Com* **40**: p. 227-256.

[9] L. Devillers, L. Vidrascu, L. Lamel, "Emotion detection in real-life spoken dialogs recorded in call center", Journal of NN, "Emotion and Brain", vol.18, 4, 407-422, may 2005.

[10] Tanja Bänziger, Klaus R. Scherer: Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus. ACII 2007: 476-487.

[11] L. Devillers, L. Vidrascu, "Emotion recognition" in the book «Speaker characterization », Christian Müller, Susanne Schötz (eds.), Springer-Verlag, (2007).

[12] H. Hardy, K, Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, C. Ursu and N. Webb: ``Multi-layer Dialog for Automated Multilingual Customer Service'', workshop ISLE, Edinburgh, dec. 2002.

[13] L. Devillers, I. Vasilescu: ``Prosodic cues for emotion characterization in real-life spoken dialogs'', Eurospeech, Geneva, septembre 2003.

[14] Boersma, P. and D. Weenink (2005). Praat: doing phonetics by computer (Version 4.3.14) [Computer program]. http://www.praat.org/.

[15] Witten, I. H. and E. Franck (2005). *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. San Francisco.

[16] L. Vidrascu, L. Devillers, Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features, *workshop Paraling07*, (2007).

[17] H. Pirker, Mixed Feelings About Using Phoneme-Level Models in Emotion Recognition, ACII 2007.

Recording audio-visual emotional databases from actors: a closer look

Carlos Busso and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering,
University of Southern California, Los Angeles, CA 90089,
busso@usc.edu, shri@sipi.usc.edu

Abstract

Research on human emotional behavior, and the development of automatic emotion recognition and animation systems, rely heavily on appropriate audio-visual databases of expressive human speech, language, gestures and postures. The use of actors to record emotional databases has been a popular approach in the study of emotions. Recently, this method has been criticized since the emotional content expressed by the actors seems to differ from the emotions observed in real-life scenarios. However, a deeper look at the current settings used in the recording of the existing corpora reveals that a key problem may not be the use of actors itself, but the ad-hoc elicitation method used in the recording. This paper discusses the main limitations of the current settings used in collecting acted emotional databases, and suggests guidelines for the design of new corpora recorded from actors that may reduce the gap observed between the laboratory condition and real-life applications. As a case study, the paper discusses the *interactive emotional dyadic motion capture database* (IEMOCAP), recently recorded at the University of Southern California (USC), which inspired the suggested guidelines.

1. Introduction

Humans use intricate orchestrations of vocal and visual modes to encode and convey intent and emotions (Busso et al., 2007b; Cowie and Cornelius, 2003; Ekman and Rosenberg, 1997). The expressive elements in the production and perception of voice, spoken language and non-verbal gestures are central to human communication. Understanding and utilizing these expressive emotional elements, hence, is key to facilitating any creative human experience, whether for learning or entertainment.

One of the major challenges in the study of emotion expression is the lack of databases with genuine interaction that comprise integrated information from the relevant communicative channels (e.g., speech, facial expression, and body posture). Human capabilities in creating expressive emotional experiences through acting provide opportunities to tackle the problem in a systematic and controlled fashion that is impossible or impractical to do with mere observational or post hoc analyses of human interaction data. Unfortunately, the current approaches used to record and post-process the emotional data obtained from actors are less than ideal to generate material that are closer to the emotions observed in real-life scenarios. The use of naïve speakers or inexperienced actors, the lack of contextualization, and the inadequate emotional descriptors are some of the main limitations found in the design of the existing emotional corpora.

The present paper considers the role of acting as a viable research methodology for studying human emotions, noting both the inherent limitations and the advantages the approach provides. This paper discusses some guidelines with the aim of designing emotional databases from actors that will closely represent the emotions observed in real-life scenarios. These guidelines, which are inspired by the lessons learned from our recent experience of recording the *interactive emotional dyadic motion capture database* (IEMOCAP) at USC, emphasizes the importance of using trained actors involved in their roles during interaction, rather than recording monologues or short sentences. Likewise, we highlight the importance of contextualization to collect genuine databases. We hope that the new generation of databases recorded from actors will decrease the discrepancy observed between laboratory and real-life conditions.

The paper is organized as follows. Section 2. presents the related work. It discusses some of the problems found in existing emotional databases. Section 3. provides suggestions that can be used to obtain more genuine emotional databases recorded from actors. As a case study, Section 4. describes the design, collection and evaluation of the IEMOCAP database, in which the suggested guidelines were followed. Finally, Section 5. gives the final remarks and our future directions.

2. Background

Acting and actors have played a key role in the study of emotions. Douglas-Cowie *et al.* reviewed some of the existing emotional databases, and concluded that in most of the corpora the subjects were asked to simulate ("act") specific emotions (Douglas-Cowie et al., 2003). In most of these cases, naïve speakers or actors without experience were asked to read short utterances or dialogs with few turns, without proper contextualization, which plays a crucial role in how we perceive (Cauldwell, 2000) and express emotions (Douglas-Cowie et al., 2005).

While desirable from the viewpoint of providing controlled elicitation, the use of actors under the current experimental settings has discarded important information observed in real-life scenarios (Douglas-Cowie et al., 2005). As a result, the performance of emotion recognition significantly degrades when automatic recognition models developed using such databases are used in real-life applications (Batliner et al., 2000; Grimm et al., 2007), where a blend of emotions is observed (Douglas-Cowie et al., 2005; Cowie et al., 2005). Differences between spontaneous ("real") and simulated ("acted") display of emotions have been studied in previous work. For example, Ekman discussed that there are certain facial action movements that subjects cannot voluntarily display when they are not experiencing certain

emotions (e.g., enjoyment, anger, fear and sadness) (Ekman, 1993). Efforts in this direction have focused on analyzing differences between real and acted smiles (Cohn and Schmidt, 2004) and eyebrow actions (Valstar et al., 2006). As a result, the research community has recently shifted to other sources of emotional databases, neglecting acting and creative arts as a viable means for studying emotions.

Examples of the most successful efforts to collect natural new emotional databases to date have been based on broadcasted television programs (Belfast naturalistic database, VAM, EmoTV) (Douglas-Cowie et al., 2003; Grimm et al., 2007; Abrilian et al., 2005), recordings in situ (lost luggage) (Scherer and Ceschi, 1997), asking subjects to recall emotional experiences (Amir et al., 2000), inducing emotion with a Wizard of Oz approach (SmartKom) (Schiel et al., 2002), using games specially designed to emotionally engage the users (EmoTaboo) (Zara et al., 2007), and inducing emotion through carefully designed human-machine interaction (SAL) (Cowie et al., 2005; Caridakis et al., 2006). However, these approaches have core limitations such as ethical issues (e.g., inducing emotions), or copyright problems that prevent the wide distribution of the corpora (Cowie et al., 2005). They are also constrained to specific domains. Furthermore, these techniques lack control over the microphone and camera locations, and the lexical and emotional content. In addition, some of the recordings have noisy visual and/or acoustic backgrounds and incomplete information from modalities (only some human communicative channels are recorded). In contrast, recording databases from actors offers the flexibility to control every aforementioned aspect.

We believe that the main problems of existing databases recorded from actors may not be the use of actors itself but the methodologies and materials used to record the existing corpora, which can be made more systematized. For example, different acting styles and methods can be utilized to enable systematic and consistent elicitation (Enos and Hirschberg, 2006). The connection with real-life ("non-acted") scenarios still needs to be clearly established. Furthermore, important aspects of human interaction such as the cognition and multimodal aspects in human interaction, largely ignored thus far, need to be carefully considered in the design of the emotional databases. When some of these aspects are included in the design of a corpus, high quality databases from actors can be recorded (Bänziger and Scherer, 2007). On the other hand, even under these limitations, it is not clear whether (and which of) these differences are clearly perceptively distinguished. For example, Schröder *et al.* reported results on perceptive experiments for induced ("real") and simulated amusement (Schröder et al., 1998). Human evaluators were asked to classify the stimuli between real and acted emotions. The results showed an average accuracy of 58%. Although the performance was over chance, the low accuracy suggests that the task was non-trivial. Human raters were not able to accurately distinguish between real and simulated emotions.

It is in this context that we pose the following question: Can specific acting methods be used to mitigate the limitations of recording emotional data from actors? The creative art forms of human acting exemplify the most enriched forms of expressive human communication. These skills have evolved over centuries, and across cultures, providing insight into how humans use their communication instruments to create and induce specific emotional percepts in others, especially in controlled and deliberate ways. The fields of theater to the contemporary cinematic arts have well-established theories and methods of pedagogy and practice of expressive communication although largely descriptive and non-quantitative offering a fertile ground for allowing research on expressive human behavior in systematic ways.

3. Guidelines to record databases from actors

The guidelines presented in this section were learned from our experience in the design, collection and evaluation of the IEMOCAP database. The main goal for this corpus was to record realistic interaction between subjects in which the emotions were naturally induced in a dialog context. We also needed to acquire detailed visual information to capture the nonverbal behavior by using motion capture technology. Under these requirements, the use of actors was the most suitable choice.

Some of the important requirements that need to be carefully considered in the design of the corpus are the actor selection, material selection, acting styles to be used, modalities to be included, and types of audiovisual sensors. The next subsections discuss some of these guidelines.

3.1. Contextualization and social setting

One of the limitations in many of the existing emotional databases is that they contain only isolated sentences or dialogs (Douglas-Cowie et al., 2003). These settings remove the discourse context, which is known to be an important component for emotion (Cauldwell, 2000). As a result, this approach challenges the actors, who have to face this task, which completely differ from the methodologies and techniques that they have been trained. Furthermore, actors are asked to read the material, which is known to differ from spontaneous speech production. Under these settings, it is not surprising that there is a gap between the expressions in such databases and the emotions observed in real scenarios. Instead of monologues and short sentences, the database should contain natural dialogues, in which the emotions are suitably and naturally elicited. Furthermore, the average duration of the dialogues should be long enough to contextualize the signs and flow of emotions (e.g., one minute (Douglas-Cowie et al., 2003)). The semantic context of the material should be congruent with the intended emotion, to avoid adding extra difficulties to the actors. The social setting is also important; having interaction between two or more actors is hypothesized to generate more authentic emotions (Douglas-Cowie et al., 2003).

One special experimental case is when the same sentences need to be spoken expressing different emotional categories. This case is important when the goal is to compare neutral versus emotional materials in terms of linguistic units (Busso and Narayanan, 2006). In this case, semantically neutral sentences need to be designed such that they can be adequately contextualized according to the intended

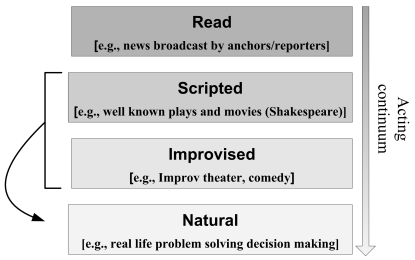


Figure 1: Acting continuum – from Fully predetermined to fully undetermined.

emotion. One approach is to record the target sentences embedded in short stories (Martin et al., 2006). For example, for the target sentence "that dress came from Asia", the following contexts could be used: sadness - the subject misses her native land; happiness - the subject receives a gift, anger - the dress was stolen.

3.2. Acting styles

The recording of the database should be as controlled as possible in terms of emotional and linguistic content. The use of specific acting methods and styles can be used to provide a systematic way to control aspects of the expressive communication forms.

Theater theory and practice, over the past several centuries, has been systematized through the work of several scholars resulting in well-known acting systems: notable examples include those of Laban, Delsarte, and Stanislavsky. It is important to investigate creativity at several points along the spectrum between fully pre-specified activity and fully improvised activity. As discussed in Section 4., acting techniques ranging from fully scripted to fully improvised can be used to balance the tradeoff between controllability and naturalness. The two most common genres of theatre are the conventional, scripted approach where dialogue and some instructions to actors are provided a priori, and *improv* in which actors are required to create much of the performance within a set of relaxed constraints (Fig. 1). While the use of scripts provides a way of constraining the semantic and emotional content of the corpus, improvisation gives the actors a considerable amount of freedom in their emotional expression. These two types of acting provide alternatives that the emotional research community should take advantage of when collecting databases (Enos and Hirschberg, 2006).

3.3. Trained actors

Unlike naïve speakers, skilled actors engaged in their role during interpersonal drama may provide a more natural representation of the emotions, avoiding exaggeration or caricature of emotions (Douglas-Cowie et al., 2003). Importantly, as the subjects display facial expressions that are closer to genuine emotions, they may start feeling the emotion, as suggested by Ekman (Ekman, 1993).

Our experiences with actors indicate that rehearsing the material in advance under the supervision of an experienced professional helps to increase the emotional quality of the data. As the actors get familiar with the material (e.g., scripts) and with their colleagues, they will be more confident during the recording.

3.4. Emotional descriptors

One of the most important aspects in the post-processing of the data is defining the emotional description that is conveyed in the data. Defining this ground reference is important since the boundaries between descriptors is usually blurred. The most common techniques to describe the emotional content of a database are discrete (category based) and continuous (primitive based) representation of the emotions. In the discrete emotional representation, categorical labels such as happiness, anger and sadness are used in a time sequence fashion. This approach has been widely used in previous work in describing human emotions. In contrast, the continuous emotional representation is based on primitive attributes. This is an alternative approach to describe the emotional content of an utterance, in which the sentences are described in terms of attributes such as valence, activation (or arousal), and dominance (Cowie and Cornelius, 2003). This approach, which has recently increased popularity in the research community, provides a more general description of the affective states of the subjects in a continuous space. Likewise, it is also useful for analyzing emotion expression variability. Both types of emotional descriptions provide complementary insights about how people display emotions. Adopting either of these schemes will depend on the research questions that will be studied.

In most of the previous emotional corpus collections, the subjects were asked to read a sentence, expressing a given emotion, which is later used as the emotional label. A drawback of this approach is that it is not guaranteed that the recorded utterances reflect the target emotions. Additionally, a given display can elicit different emotional percepts. To avoid these problems, the emotional description should rely on perceptual human evaluation collected from as many evaluators as possible ($\geq$3).

Subjective evaluations are expensive and, therefore, need to be suitably designed in advance (running pilot tests is highly suggested). When categorical description is adopted, a critical aspect is the emotional classes included in the evaluation. On the one hand, if the number of emotions is too extensive, the agreement between evaluators will be low (which is an inherent problem of emotional subjectivity). Ad-hoc solutions such as clustering *similar* emotional categories after the perceptive evaluation should be avoided, since the emotional partition will most probably differ from the one obtained with the new labels. On the other hand, if the list of emotions is limited, the emotional description of the utterances will be poor and likely less accurate.

Another important aspect is the order in which the material will be presented. We suggest presenting the material in order (i.e., not in isolated fashion), so that the evaluators can judge the emotional content based on the sequential development of the dialogs. Likewise, all available modalities should be presented, so the evaluators can use all their senses to assess the emotion. Finally, long and tedious subjective evaluations should be avoided.

Figure 2: VICON motion capture system with 8 cameras, and an actress showing the markers on the face and headband.

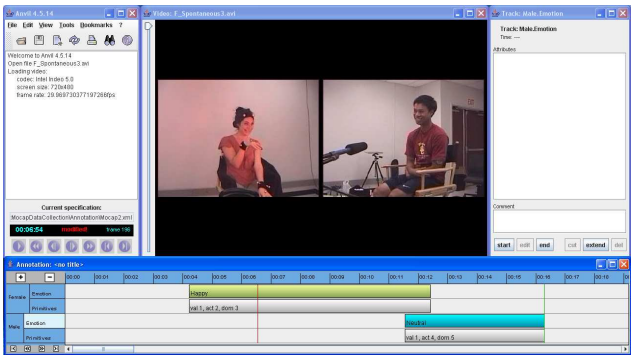


Figure 3: ANVIL annotation tool used for emotion evaluation. The tool is set for categorical and primitive based evaluation.

4. A case study: The USC IEMOCAP corpus

As a case study, we recently collected an audiovisual database, we refer to as the *interactive emotional dyadic motion capture database* (IEMOCAP) (Busso et al., 2007a). This is an extensive corpus with over twelve hours of data comprising multimodal information. This corpus was designed, collected and evaluated following the guidelines suggested in Section 3.. This section briefly describes this multimodal corpus.

4.1. Designing the corpus

This database was collected from seven professional actors and three senior students (5 female and 5 male) from the Drama Department at USC. These experienced actors were recorded in dyadic sessions to facilitate a social setting suitable for natural interaction (Sec. 3.1.).

In contrast to providing material that an actor reads under the target emotional condition, the two approaches mentioned in Section 3.2. were selected: the use of plays (scripted sessions), and improvisation-based hypothetical scenarios (spontaneous sessions). The first approach is based on a set of scripts that the subjects were asked to memorize and rehearse. In the second approach, the subjects were asked to improvise based on hypothetical scenarios that were designed to elicit specific emotions (happiness, anger, sadness, frustration and neutral states). These recording settings are familiar to the actors since they were trained to memorize and improvise scripts (Sec. 3.3.).

4.2. Data collection

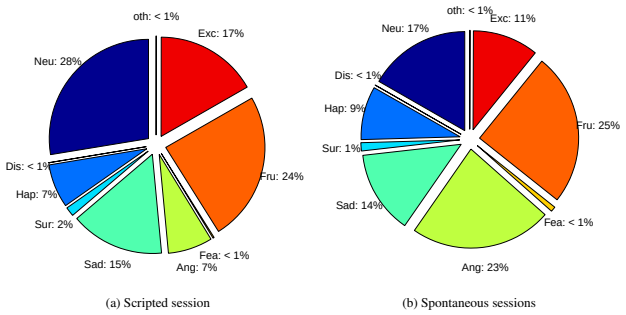
To capture non-verbal behavior of the subjects, markers were attached to their face, head and hands, which provide detailed information about their facial expression and hand movements. To track those markers, a VICON motion capture system with 8 cameras was used (Fig. 2). Due to equipment constraints, only one of the subjects' movements were captured at a time. Then, the markers were placed on the other subject and the material was recorded again.

4.3. Evaluation

After the data were recorded, the dialogs were manually segmented at the dialog turn level. The emotional categorical labels in this corpus were assigned based on agreements derived from subjective emotional evaluations (3 evaluators per sentence). For the aforementioned purpose, the *annotation of video and spoken language* tool ANVIL (Kipp, 2001) was used (Fig. 3). This tool is particularly useful to jointly annotate verbal and nonverbal behaviors observed from the actors. The evaluators were asked to sequentially assess the turns, after watching the videos. Thus, the acoustic and visual channels, and the previous turns in the dialog were available for the emotional assessment, so that the evaluators could judge the emotional content based on the sequential development of the dialogs (Sec. 3.4.). For the evaluation, the emotional categories surprise, fear, disgust, excited, and other were also included. While categorical description is already evaluated, we are currently assessing the database in terms of the primitive attributes to have a more complete emotional description.

Majority voting was used to tag the sentences with the emotional categories. Under this criterion, 74.6% of the sentences were assigned one emotional category (spontaneous sessions: 83.1%; scripted session: 66.9%). For the sentences in which the evaluators reached agreement, the resulting Fleiss kappa statistic was κ= 0.40 (spontaneous sessions: κ = 0.43; scripted session: κ = 0.36). These levels of agreement, which are considered as fair/moderate agreement, are expected since people have different perception and interpretation of the emotions. They also show the difficulties in the assignment of emotional labels. Interestingly, the results reveal that for the spontaneous sessions the levels of inter-evaluator agreement are higher than in the scripted sessions. While spontaneous sessions were designed to target specific emotions, portions of the scripted sessions elicited a wider range of emotion categories, increasing the confusion between evaluators.

Figure 4 shows the emotional content of the database. For the target emotions (happiness, anger, sadness, frustration and neutral states), the figure indicates that a balanced emotional content was obtained which validates the proposed design. Notice that for scripted sessions, the emotional labels are less balanced than for spontaneous sessions. From a design viewpoint, these results suggest that improvisation may be preferred if a balanced corpus with higher agreement level is required.



(a) Scripted session (b) Spontaneous sessions

Figure 4: Distribution of the data for each emotional category.

5. Conclusions

This paper described guidelines with the aim of designing controlled emotional databases from actors that are closer to the emotions observed in real-life scenarios. Based on the limitations of current settings used in the recording of existing corpora, we discussed the importance of contextualization, the use of skilled actors, the use of different acting styles and suitable emotional descriptors.

As a case study, the paper presented the IEMOCAP database. Based on the settings used to elicit the emotions and the achieved results, we consider that the emotional quality of this database is closer to natural than those from prior elicitation settings. The quality of this database suggests that genuine realization of the emotions can be recorded from actors when the settings are carefully designed.

While this methodology was a significant step forward in the use of actors in emotions in research, it did not exploit or control for the nature of the expressive behavior such as through specific acting styles or the nature of improvisation. Further analysis is needed to identify the recording methodologies that will aid emotional recording from actors that resemble real emotions observed in daily human interaction. In fact, acting methods such as the one proposed by Stanislavsky (Carnicke, 1998), in which the actors are encouraged to feel their characters, could be exploited to capture realistic realization of the emotions. These are some of the goals of our future work.

Acknowledgements

This research was supported in part by funds from the NSF, and the Department of the Army. The authors wish to thanks the anonymous reviewers for their thoughtful and insightful comments. Thank also go to the colleagues in the emotion research group for their valuable comments.

6. References

S. Abrilian, L. Devillers, S. Buisine, and J.C.Martin. 2005. EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *11th International Conference on Human-Computer Interaction (HCI 2005)*, pages 195–200, Las Vegas, Nevada, USA, July.

N. Amir, S. Ron, and N. Laor. 2000. Analysis of an emotional speech corpus in Hebrew based on objective criteria. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 29–33, Newcastle, Northern Ireland, UK, September.

T. Bänziger and K.R. Scherer. 2007. Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In A. Paiva, R. Prada, and R.W. Picard, editors, *Affective Computing and Intelligent Interaction (ACII 2007), Lecture Notes in Artificial Intelligence 4738*, pages 476–487. Springer-Verlag Press, Berlin, Germany, September.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. 2000. Desperately seeking emotions or: actors, wizards and human beings. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 195–200, Newcastle, Northern Ireland, UK, September.

C. Busso and S.S. Narayanan. 2006. Interplay between linguistic and affective goals in facial expression during emotional utterances. In *7th International Seminar on Speech Production (ISSP 2006)*, pages 549–556, Ubatuba-SP, Brazil, December.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2007a. IEMOCAP: Interactive emotional dyadic motion capture database. *Submitted to Journal of Language Resources and Evaluation*.

C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. 2007b. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086, March.

G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis. 2006. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the 8th international conference on Multimodal interfaces (ICMI 2006)*, pages 146–154, Banff, Alberta, Canada, November.

S.M. Carnicke. 1998. *Stanislavsky in focus*. Routledge, Taylor & Francis Group, Oxford, UK.

R. Cauldwell. 2000. Where did the anger go? the role of context in interpreting emotion in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 127–131, Newcastle, Northern Ireland, UK, September.

J. Cohn and K. Schmidt. 2004. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:1–12, March.

R. Cowie and R.R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, April.

R. Cowie, E. Douglas-Cowie, and C. Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4):371–388, May.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, April.

E. Douglas-Cowie, L. Devillers, J.C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. 2005. Multimodal

databases of everyday emotion: Facing up to complexity. In *9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pages 813–816, Lisbon, Portugal, September.

P. Ekman and E.L. Rosenberg. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. Oxford University Press, New York, NY, USA.

P. Ekman. 1993. Facial expression and emotion. *American Psychologist*, 48(4):384–392, April.

F. Enos and J. Hirschberg. 2006. A framework for eliciting emotional speech: Capitalizing on the actors process. In *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, pages 6–10, Genoa,Italy, May.

M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800, October-November.

M. Kipp. 2001. ANVIL - a generic annotation tool for multimodal dialogue. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, Aalborg, Denmark, September.

O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE'05 audio-visual emotion database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW 2006)*, Atlanta, GA, USA, April.

K.R. Scherer and G. Ceschi. 1997. Lost luggage: A field study of emotionantecedent appraisal. *Motivation and Emotion*, 21(3):211–235, September.

F. Schiel, S. Steininger, and U. Türk. 2002. The SmartKom multimodal corpus at BAS. In *Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain, May.

M. Schröder, V. Aubergé, and M.A. Cathiard. 1998. Can we hear smiles? In *5th International Conference on Spoken Language Processing(ICSLP 1998)*, pages 559–562, Sydney,Australia, November-December.

M.F. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn. 2006. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces (ICMI 2006)*, pages 162–170, November.

A. Zara, V. Maffiolo, J.C. Martin, and L. Devillers. 2007. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In A. Paiva, R. Prada, and R.W. Picard, editors, *Affective Computing and Intelligent Interaction (ACII 2007), Lecture Notes in Artificial Intelligence 4738*, pages 464–475. Springer-Verlag Press, Berlin, Germany, September.

Acted vs. spontaneous expressive speech: perception with inter-individual variability

Nicolas Audibert[1], Véronique Aubergé[1] and Albert Rilliard[2]

[1] Gipsa-lab Speech & Cognition Dept (Institut de la Communication Parlée),

CNRS UMR 5216/Université Stendhal, 38040 Grenoble Cedex 9, France

[2] LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

E-mail: {Nicolas.Audibert, Veronique.Auberge}@gipsa-lab.inpg.fr; Albert.Rilliard@limsi.fr

Abstract

This paper reports how acted vs. spontaneous expressive speech can be discriminated by human listeners, with various performances across listeners (in line with preliminary results for amusement by Aubergé et al. (2003)). The speech material was taken from the Sound Teacher/E-Wiz corpus (Aubergé et al., 2004), for 4 French-speaking actors trapped in spontaneous expressive monoword utterances, and acting immediately after in an acting protocol supposed to optimal for them. Pairs of acted vs. spontaneous stimuli, expressing affective states related to anxiety, irritation and satisfaction, were rated by 33 native French listeners in audio-only, visual-only and audiovisual conditions. In visual-only condition, 70% of listeners were able to discriminate acted vs. spontaneous pairs over chance level, for 78% in audio-only condition and up to 85% in audio-visual condition. A strong listener effect confirms the hypothesis of a variable affective competence for separating involuntary vs. simulated affects (Aubergé, 2002). One feature used by listeners in the acoustic task of discrimination can be the perceived emotional intensity, in accordance with the measurement of this intensity level for the same stimuli from a previous perception experiment by Laukka et al. (2007).

1. Introduction

Although the question of the validity of corpora of acted emotional speech for the modeling of affective speech has been debated (Campbell, 2000), leading to an increase of the research effort directed towards spontaneous emotional speech (see for instance Batliner et al. (2004)), few studies have been comparing the performances of acted vs. spontaneous speech to our knowledge. Aubergé et al. (2003) proposed that acted vs. spontaneous amusement could be discriminated, judges discrimination competences being highly variable independently of the speaker's acting skills. More recently, Wilting et al. (2006) recorded naïve Dutch participants without particular acting skills while inducting positive and negative moods using the Velten procedure, prior to asking them to produce the same utterances while simulating similar moods. Though acted and spontaneous utterances were not directly compared, a perception experiment in visual condition showed that acted expressions were rated as more intense than spontaneous ones.

Such findings are coherent with a strong hypothesis that we proposed for the processing of affects (after Fonagy (1984) and Scherer (2001)): affects are cognitively distinguished following two ways of control by the speaker: voluntary vs. involuntary, and not as a function of the affective information carried by the expressions. In this view (Aubergé, 2002), authentic emotions are performed by involuntary control (the "push effect" in Scherer's model (2001)). The speaker is also able to reproduce the expressions of past emotions outside the body loop described by Damasio (2003) and through a voluntary control, that are the social affects in the hypothesis we defend. This implies that a same value of affect can be processed voluntarily or not. The voluntary performance, that informs the speech acts generated by the speaker, can be very spontaneous and sincere, but it does not reflect the same processing as an authentic emotional expression. We claim (Aubergé, 2002) that this competence, central in the communication processing and producing the largest part of expressivity in interactions, is not anchored in the same timing processing (voluntary/social expressions would be anchored in the time of the language organization). Over the basic reproduction of emotions, cultures and languages have developed large specific scales of such voluntarily controlled affect, that we call attitudes.

We make the hypothesis that this competence of reproducing expressions is used by the actors on speech acts given by the scenario, especially when these actors belong to an acting field (1) devoted to simulate to be very authentic in the given acting story context (2) based on method using the memory of previously felt emotions. That is precisely the field of the actors participating to this experiment.

The main question asked in the present experiment is whether expressions with same emotional values through voluntary vs. involuntary control, i.e. in this case through a simulation by acting vs. involuntary felt emotions, can be discriminated by human listeners, and whether all humans have a similar competence for accessing these cues.

Moreover Aubergé et al (2003) showed that the acoustic information is integrated to the visual decoding of affective values, even when the face carries strong affective information. More generally it has been shown that emotional expressions must be considered as multi-modal processing (e.g. Scherer & Ellgring, 2007). The study presented in this paper thus focuses on multimodal expressions of affective speech, trying to

separate the information carried by different modalities, even though the face also carries information about speech and this information consequently cannot be considered as additive across modalities.

2. Acted vs. spontaneous speech collection

The French expressive corpus E-Wiz (Aubergé et al., 2004) was recorded using the Wizard of Oz technique, in which the subject is convinced to be interacting with a complex person-machine interface while the apparent behavior of the application is remote-controlled by the wizard. Subjects were asked to participate in the testing prior to its commercialization of a so-called voice-recognition-based language-learning software. In this task the subjects had to interact with the system using a command language composed of the French monosyllabic color names [bʁik], [ʒon], [ʁuʒ], [sabl] and [vɛʁ] and the command [paʒsʮivãt] (*next page*). The performances of the 17 subjects participating in the experiment were manipulated to induce positive then negative emotions, and the affects expressed were labeled by the subjects themselves from the video recording, as a first labeling step before perceptive validation. A particular protocol was set up for the 7 subjects who were also actors: those subjects were requested immediately after the Wizard of Oz task to produce again the affects they reported to have felt during the experiment on the same utterances as well as the most frequently studied emotions (sadness, anger, fear, disgust, surprise and joy), using their acting methods. The experimenters insisted that the actors should express the affects felt in the experiment the same way they had been feeling them just before. The actors recruited for this task were practicing improvisation theater and/or street acting, and used past felt emotions as a basis for expressing emotions, as described in (Enos & Hirschberg, 2006). All of them reported the experimental set-up as optimal for being in good acting conditions with regards to their acting habits.
A first experiment using both acted and spontaneous utterances from the E-Wiz corpus in audio-only condition, and focusing on the typicality of vocal expressions of emotion, was conducted by Laukka et al (2007). In this experiment, 47 acted and 146 spontaneous utterances produced by 6 actors (3 males, 3 females) were validated and rated for emotional intensity in a pre-test, showing a higher perceived emotional intensity for acted utterances vs. spontaneous ones, in line with results obtained by Wilting et al. (2006) in visual-only condition. We present in this paper the results of a discrimination task between acted and spontaneous utterances, based on productions of the speakers evaluated in this pre-test.

3. Experimental protocol

A subset of stimuli matched between acted and spontaneous expressions and suitable for being evaluated in a perceptive discrimination task was selected. The criteria for selection were that paired stimuli should be of equal length, produced by the same speaker, and previously rated with comparable emotional intensities in audio condition, according to the measurement made in (Laukka et al., 2007). As a result, 24 pairs of acted vs. spontaneous stimuli produced by the 4 actors (2 male and 2 female) showing the more expressiveness in their acted productions were retained.
The selected pairs were presented to subjects with a latency of 1.5 seconds between both, with 3 presentation conditions: audio only (A), visual only (V) and audiovisual (AV). Stimuli were presented grouped by condition and randomly sorted within each condition, AV condition being always the last one while A and V conditions were alternatively chosen as first condition. Each pair was presented twice in each condition with the spontaneous utterance in first or second position to compensate for a possible effect of the presentation order. After each presentation of a pair the subject was requested to indicate which stimulus he considered to be the spontaneous one, using a slider ranging from 'certainly the first one' to 'certainly the second one', which initial position was set to the middle. This slider was intended to capture both discrimination and confidence level, similarly to the procedure developed by Bänziger (2003). Answers could be validated only after the slider had been moved. The presentation of stimuli and the recording of subjects' answers were automated through a user interface developed with the Revolution software for the needs of the experiment. Subjects were explained the main goals of the experiment as well as the context of the corpus recording prior to the actual beginning of the task. 33 native French subjects (15 male, 18 female, mean age 33.1) without known hearing problems took the listening test, which lasted 25 minutes in average.

4. Analysis

4.1 Statistical method

Slider position values were converted into discrimination scores (right vs. wrong answers) according to the direction in which the slider had been moved, and into a confidence level according to the distance from the slider position to the initial position. The mean discrimination score for each pair was only moderately correlated to the confidence level (r=.408 in audio-only condition, r=.690 in visual-only condition, r=.583 in audiovisual condition, r=.622 overall). Discrimination scores for different presentation conditions and speakers are summarized in figure 1.
Discrimination scores and confidence levels were analyzed using repeated measures analyses of variance (ANOVA) with listener, speaker, emotion class, presentation condition, utterance length and presentation order as fixed factors.
As most of significant effects were found to be the same on discrimination and confidence scores, the present study mainly focuses on discrimination and reports only a few remarkable effects on confidence scores.

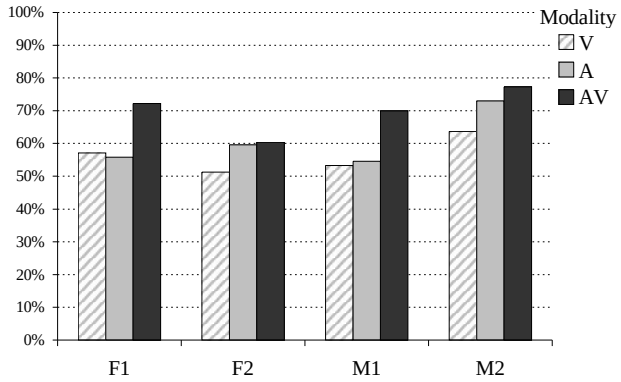


Figure 1: Overall results according to speakers and presentation conditions

4.2 Differences in listeners' performances

A strong listener effect was found on discrimination of spontaneous vs. acted utterances ($F(1,31)=801.58$, $p<.001$), in line with results of Aubergé et al. (2003) on amusement. As a matter of fact, discrimination scores of different listeners range from 32.7% to 80.6% of correctly classified pairs. Though the listener effect on the rated confidence level was also highly significant ($F(1,31)=220.23$, $p<.001$), strong conclusions should not be drawn from this result as it might more reflect different strategies in the use of the slider than differences in subjects' abilities.

Although listeners' competences for discriminating acted vs. spontaneous expressions were highly variable, they did not appear to show preferences for particular speakers independently of their acting skills. As a matter of fact, Cronbach's alpha value for individual discrimination performances on different speakers' production was quite high (alpha=0.8671), indicating that listeners competences were consistent across different speakers.

Figure 2 presents the distribution of listeners' discrimination scores for each presentation condition and overall. As it can be observed from this chart, 70% of listeners were able to correctly discriminate more than half of the presented pairs in visual-only condition, while 79% did in audio–only condition and 85% did in audiovisual condition (85% overall).

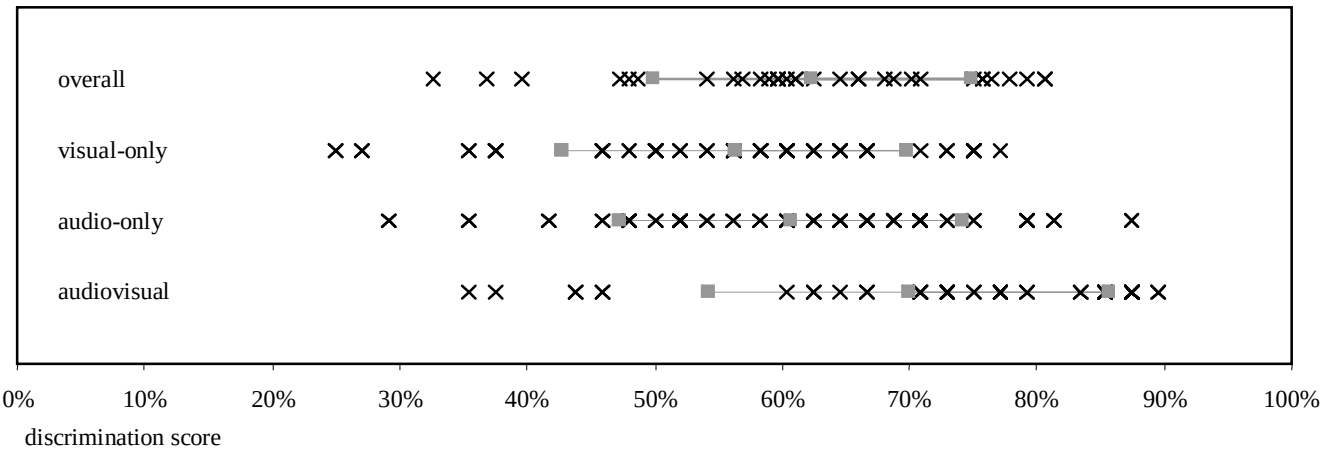


Figure 2: Distribution of overall and per presentation condition listeners' discrimination scores. Grey boxes and solid lines indicate mean and standard deviation for each condition

4.3. Discrimination across modalities

A significant main effect of the presentation condition on discrimination scores was found ($F(2,62)=21.33$, $p<.001$). Contrasts between conditions show a significant gain of discrimination for the audiovisual condition when compared to the audio-only and visual-only ($p<.001$ for both contrasts) conditions, while the difference in discrimination scores between audio-only and visual only conditions was non-significant. However this advantage of audiovisual condition against audio-only and visual-only conditions was not constant across different speakers, as illustrated by figure 1: the effect of condition was indeed non-significant for speaker F2, and the discrimination increase from audio-only to audiovisual condition was non-significant for speaker M2.

4.4. Speaker effect

A significant main effect of the speaker ($F(3,93)=16.05$, $p<.001$) was also observed. Only spontaneous utterances of speaker M2 were significantly better discriminated than those of all other speakers ($p<.001$ for all 3 contrasts), indicating that this speaker was less successful than the other actors in pretending that he was actually expressing spontaneous affects. On the other hand, all 3 other speakers' productions were discriminated with similar scores, although a large part of listeners reported to have considered the discrimination task as more difficult for speakers F2 and M1 than for the two other speakers. This intuition of listeners was illustrated by the fact that, whereas speaker M2 also received the highest confidence ratings, confidence ratings attributed to utterances of speaker F1 were significantly higher than those of speakers F2 and M1 ($p<.001$ for both contrasts). This

speaker effect was stronger in audio-only and audiovisual conditions ($p<.001$ for both) than in visual-only condition ($p<.05$).

4.5. Other effects

No overall effect of the emotion class was found, suggesting that subjects have similar abilities for discriminating a spontaneous vs. an acted expression whatever the emotion expressed. The effect of length was just significant, with utterances of [paʒsʮivãt] slightly better discriminated than monosyllabic utterances ($F(1,1)=6.33$, $p<.05$).

The order of presentation of the stimuli in the pairs (spontaneous then acted vs. acted then spontaneous) was globally significant ($F(1,1)=8.32$, $p<.01$), with a higher discrimination score for pairs in which the spontaneous stimulus was presented first. This effect is however only significant in the visual only condition ($p<.001$), and related to the speakers particular production: the effect is indeed significant ($p<.01$) for only two of them (speaker M2 who was the best recognized, i.e. the less good actor, and speaker F2 for whom the discrimination was the lowest, though not significantly different from other actors), which are also those for whom audiovisual discrimination scores were not significantly better when compared to audio-only. We did not yet proceed to a complete objective analysis of the visual expressions, but observable gestures amplitudes of those two actors are obviously larger than those of the two others. Moreover both of them moved almost systematically the head downwards in spontaneous speech. A possible explanation for this presentation order effect could be ecological strongly informative reference given by the spontaneous stimulus presented first, for a better discriminative comparison with the second, acted, stimulus.

Though both discrimination and confidence were higher for female than for male listeners, especially in audio-only condition, those differences were found to be non-significant. Although statistical significance cannot be calculated in that case, a particular result is worth being noted: for two male subjects performing better than the average in audio-only condition but worse than the average on visual condition, the visual information seem to have largely lowered discrimination performances in audiovisual condition. The discrimination score of those two subjects was indeed more than 20% lower in audiovisual condition than in audio-only condition.

Correlations between duration differences of the spontaneous vs. acted stimuli presented in each pair (ranging from -480 to 760 ms) and averaged discrimination and confidence scores were calculated in order to look for a possible account of this difference in the discrimination performances. However those correlations were very low ($r=.047$ for discrimination, $r=.037$ for confidence), suggesting that this information was not used as a cue for discrimination.

4.6. Emotional intensities and discrimination

The partial correlations for different presentation conditions between discrimination scores or confidence ratings, and differences of perceived emotional intensities in the pair from Laukka et al. (2007) are presented in table 2 with the overall correlations. The number of pairs for which those correlations can be calculated does not allow to draw strong conclusions from these values. However the stronger correlation in audio-only condition, especially between discrimination scores and perceived intensity differences, suggests that the difference in perceived emotional intensity may at least partly account for the discrimination between spontaneous and acted utterances. As emotional intensity ratings were given from presentations in audio-only condition, it is not a surprising result to find higher correlations for this condition.

condition	**A**	**V**	**AV**	**overall**
discrimination	r=.745	r=.131	r=.335	r=.415
confidence	r=.402	r=.147	r=.283	r=.250

Table 1: Correlations between difference in perceived emotional intensity in the pair (extracted from 0) and discrimination scores or confidence level

Although pairs with the highest difference in perceived emotional intensity appear to be among the best discriminated pairs in A condition, suggesting that perceived emotional intensity might be a strong cue for discrimination when accessible, listeners did not rely only on that feature for discriminating spontaneous vs. acted emotions.

As a matter of fact, among the 3 pairs for which the difference in emotional intensity between spontaneous and acted was the weakest, ranging from -4.6% to 4,6%, only the expressions of irritation produced by speaker M1 on monosyllabic utterances were poorly discriminated (correctly discriminated by only 42,4% of listeners in audio-only condition) while the expressions of irritation of speaker F1 on [paʒsʮivãt], evaluated with the same emotional intensity, were correctly discriminated by 62,1% of listeners. On the other hand expressions of irritation on [paʒsʮivãt] by speaker F2, for which the acted expression had been evaluated as 16.6% more intense than the spontaneous one, were among the most poorly discriminated in audio-only condition (correctly discriminated by only 40,9% of listeners).

5. Conclusion

The results presented in this paper suggest that listeners are globally able to discriminate acted vs. spontaneous multimodal expressions, without an effect of the emotion (the only three kinds of emotional information evaluated are indeed quite balanced in terms of activation and valence), and with a strong listener effect.

The perceived emotional intensity, previously measured between acted and spontaneous in auditory condition by Laukka et al. (2007), and in visual condition by Wilting et al. (2006) on different data, might be an artifact explaining part of the discrimination scores. Such variation in perceived emotional intensities can be considered as a major bias in this kind of discrimination tasks, as pointed out by Aubergé et al. (2003). However differences in perceived emotional intensity can definitely not account for the whole variability. In order to evaluate more systematically to what extent perceived emotional intensities can be linked to discrimination of acted vs. spontaneous expressions, a perception experiment using the same set of 48 stimuli presented in the same conditions is currently being prepared.

Even if the chosen actors were certainly not among the ones recognized as the best, three of them out of four could trick the less competent listeners, and are typically the kind of actors frequently chosen for recording emotional databases. The acted speech, commonly used as a reference for studies on involuntary emotions, could be considered carefully knowing the human discrimination ability. Over such ability to discriminate, the question of the variability of human competence (that may be related to the affective quotient) for identifying simulation, that is in our proposals the modal processing used in interaction by a speaker expressing his attitude whatever his sincerity (including the reproduction of involuntary emotions), remains an open question that would deserve being specifically studied.

The acoustic and visual analysis of the stimuli, according to the perceptive results, is under progress, on the basis of a strong hypothesis (Aubergé, 2002) on the difference of timing anchoring between involuntary and voluntary (social) emotions.

Since the cognitive processing of acted speech cannot be directly related to the cognition of voluntary expressed emotions, i.e. the social affects, a further experiment will be to catch and perceptively compare spontaneous emotions vs. spontaneous attitudes (reduced to emotion values).

6. Acknowledgments

The authors warmly thank Christophe Savariaux for technical support, and Petri Laukka for initiating the project. We also thank the actors and the listeners who kindly participated in the experiment.

7. References

Aubergé, V., Audibert, N., Rilliard, A. (2004). E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 179-182.

Aubergé V. (2002). A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. In *Proceedings of the 1st International Conference on Speech Prosody*, Aix-en-Provence, France, pp. 151-155.

Aubergé, V., Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication* 40, *Special issue on Emotional Speech*, pp. 87-97.

Bänziger, T. (2004). *Communication vocale des émotions. Perception de l'expression vocale et attributions émotionnelles*. PhD thesis, University of Geneva.

Batliner, A.; Hacker, C.; Steidl, S.; Nöth, E.; D'Arcy, S.; Russel, M.; Wong, M., 2004. 'You stupid tin box' - children interacting with the AIBO robot: Across-linguistic emotional speech corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 171–174.

Campbell, N., 2000. Databases of Emotional Speech. In *Proceedings of the 1st ISCA Workshop on Speech and Emotions*, Newcastle, North Ireland, pp. 34-38.

Enos, F.; Hirschberg, J., 2006. A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process. In *Proceedings of the 1st International Workshop on Corpora for Research on Emotion and Affect*, Genova, Italy, pp. 6-10.

Damasio A. R. (2003). *Looking for Spinoza. Joy, Sorrow an the Feeling Brain.* Orlando:FL/Harcourt.

Fonagy, I. (1983). *La vive voix.* Paris:Payot.

Laukka, P., Audibert, N., Aubergé, V. (2007). Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"? In *Proceedings of the 1st International Workshop on Paralinguistic Speech - between models and data*, Saarbrücken, Germany, pp. 1-4.

Scherer, KR, Ellgring, H. (2007). Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion*, 7(1), pp. 158-171.

Scherer K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In Scherer K. R., Schorr A., & Johnstone T. (eds.), *Appraisal processes in emotion: Theory, Methods, Research*, Oxford University Press, pp. 92-120.

Wilting, J., Krahmer, E., Swerts, M. (2006). Real vs. acted emotional speech. In *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, PA, USA (CD-ROM proceedings).

Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus

A. Batliner, S. Steidl, E. Nöth

Chair of Pattern Recognition, Friedrich-Alexander-University Erlangen-Nuremberg (FAU), Germany
email: {batliner,steidl,noeth}@informatik.uni-erlangen.de

Abstract

We report on a thoroughly processed and annotated German emotional speech database (children interacting with Sony's Aibo robot): 51 children, some 48 k words, 9.2 hours of speech, 5 labellers, word-based annotation of emotional user states. Several additional annotations as well as a mapping onto higher units of different granularity have been carried out. The database will eventually be made available for scientific use; in the licensing agreement, we plan to include mandatory benchmark constellations in order to make a comparison across sites possible.

1. Introduction[1]

Even if the terminology has not been standardised yet – there is no agreement as for the *exact* meaning of 'naturalistic', 'realistic', 'spontaneous', etc. – it is generally agreed upon that non-acted emotional databases should be aimed at. This might not be mandatory for generic, basic research but holds especially if we think of any application that eventually, outside of the laboratory, has to deal with non-acted data – simply because classifiers have to be trained with data that are as close as possible to the 'real' data. Obviously, the effort needed for designing, recording, and annotating spontaneous emotional databases is way higher than the one needed for acted data; thus, some acted databases are (freely) available such as the Berlin Database of Emotional Speech 'Emo-DB' (Burkhardt et al., 2005) or the Danish Emotional Speech Database 'DES' (Engberg et al., 1997) but, at least to our knowledge, no spontaneous one, at least not on similar conditions. Moreover, privacy reasons often prevent such data to be released to third parties. Thus, access to spontaneous data is the most severe bottleneck for a 'realistic' processing and emotion classification.

In the years 2002-2004, we have collected and processed a spontaneous emotional German database at FAU Erlangen within the EU-project PF-STAR. In the years 2005-2007, this database has been further annotated, and processed outside and within the so-called CEICES initiative (Batliner et al., 2006) within the NoE HUMAINE. There exist several publications on experiments using this database; however, its description has always been rather short, concentrating on those aspects that we focused on in the respective papers; this was simply due to the usual space restriction. As we eventually decided to release the database for scientific use, in this paper we want to give a condensed overview of aspects, annotations, and conditions of use. After a general description of the design and the recordings, we will give an account of the annotations which will be made available. In the end, we decided to call the database the 'FAU Aibo Emotion Corpus' (abbr.: FAU Aibo) because there exist other 'Aibo' corpora with emotional speech, cf. (Tato et al., 2002; Küstner et al., 2004).

Of course, we do not conceive the strategies chosen and presented in this paper as the only and best ones but as reasonable choices. At the end of most (sub-)sections, we will motivate our choices and partly discuss them against the backdrop of possible alternative solutions. These comments will be given in italics.

2. Material

The general frame for FAU Aibo is human–robot communication, children's speech, and the elicitation and subsequent recognition of emotional user states. The robot is Sony's (dog-like) robot AIBO. The basic idea is to combine a rather so far neglected type of corpus (children's speech) with 'natural' emotional speech within a Wizard-of-Oz task. The children were not told to use specific instructions but to talk to the Aibo like they would talk to a friend. They were led to believe that the Aibo is responding to their commands, but the robot is actually being controlled by a human operator, using the 'Aibo Navigator' software over a wireless LAN (the existing Aibo speech recognition module is not used). The wizard causes the Aibo to perform a fixed, pre-determined sequence of actions, which takes no account of what the child says. For the sequence of Aibo's actions, we tried to find a good compromise between obedient and disobedient behaviour: we wanted to provoke the children in order to elicit emotional behaviour but of course we did not want to run the risk that they break off the experiment. The children believed that the Aibo was reacting to their orders - albeit often not immediately. In fact, it was the other way round: the Aibo always strictly followed the same screen-plot, and the children had to align their orders to it's actions.

The data was collected from 51 children (age 10 - 13, 21 male, 30 female). The children were from two different schools, Mont and Ohm; the recordings took place in the resp. class-rooms. Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals is 48 kHz, quantisation is 16 bit. The data is downsampled to 16 kHz. Each record-

[1]This work was funded by the EU in the projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-50742. The database has been further processed in the initiative CEICES within HUMAINE (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates). The responsibility for the contents of this study lies with the authors.

ing session took some 30 minutes. The speech data were segmented automatically into speech files ('turns'), triggering a turn boundary at pauses $\geq$ 1.5 seconds. Note that here, the term 'turn' does not imply any linguistic meaning; however, it turned out that only in very few cases, this criterion wrongly decided in favour of a turn boundary instead of (implicitly) modelling a hesitation pause. Because of the experimental setup, these recordings contain a huge amount of silence (reaction time of the Aibo), which caused a noticeable reduction of recorded speech after raw segmentation; finally we obtained about 9.2 hours of speech.

Children as early adapters within an edu-/entertainment scenario should be plausible addressees for automatic emotion modelling. Sometimes it has been doubted that they are 'representative' because they might behave unlike adults. Of course, speech recognition and feature extraction have to be adapted slightly – the same way as procedures designed only for male speakers have to be adapted for female speakers. Of course, the children can display group-specific tendencies but there is no indication that they behaved differently from adults in a fundamental way.

3. Further Processing

3.1. Transliteration and word lexica

The orthographic transliteration – in pared down VERBMOBIL notation – was done by advanced students and cross-checked by the supervisor. The phonetic word lexicon is in SAMPA notation. In addition, we established a syntactic-semantic word lexicon, with coarse part-of-speech (POS) labels per word (six classes), and with coarse higher semantic labels per word (six classes modelling valence and some other word types such as vocative).

While modelling linguistic information, normally words are not used as such but processed somehow – at least they are stemmatised for, e.g., bag-of-word modelling. In our experience, such a very coarse mapping onto six POS or higher semantic classes only still yields a pretty good classification performance on the spoken word chain. Of course, there are many other mappings which can be conceived. However, our 'simple' classes can be used as benchmark for other approaches, cf. below.

3.2. Word-based emotion annotation

In other studies, the unit of analysis is normally given trivially – a read sentence, a dialogue move, etc. – or defined intuitively. We conceive the word as the smallest possible emotional unit; even if we cannot exclude the possibility of changing emotions within the same word, this will definitely be a rather exotic exception. By annotating word-based, we are later on free to map words onto longer units. Our strategy thus allows to find 'optimal' units of analysis on an empirical basis.

Five labellers (advanced students of linguistics, 4 females, 1 male) listened to the speech files in sequential order and annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes which were obtained by inspection of the data; we do not claim that they represent children's emotions in general, only that they are adequate for the modelling of the behaviour of these children in this specific scenario. We resort to majority voting (henceforth MV): if three or more labellers agree, the label is attributed to the word; if four or five labellers agree, we assume some sort of prototypes. The following raw labels were used; in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i. e., irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral, but not belonging to the other categories (3), *neutral* (39169); 4707 words had no MV; all in all, there were 48401 words. *joyful* and *angry* belong to the 'big' emotions, the other ones rather to 'emotion-related/emotion-prone' user states. The state *emphatic* has been introduced because it can be seen as a possible indication of some (starting) trouble in communication and by that, as a sort of 'pre-emotional' state (Batliner et al., 2005; Batliner et al., 2008).

A single database is no omnibus in the sense that choosing a specific scenario for the recordings pre-defines the range of classes one can observe; what cannot be observed cannot be modelled. However, we claim that our data are fairly representative for realistic data: only a few of the 'classic', big n emotions, and a very skewed distribution. Instead, one 'emotion-related' state comes on the scene, i. e. motherese. Emphatic is, in fact, just a possible prestage of emotion. However, in some form it will often be observed in such realistic settings; thus it makes sense to model it.

3.3. Partitioning into sub-samples

Some of the labels are very sparse; if we only take labels with more than 50 MVs, this 7-class problem is most interesting from a methodological point of view, cf. the new dimensional representation of these seven category labels in (Batliner et al., 2008). However, the distribution of classes is very unbalanced. Therefore, we downsampled *neutral* and *emphatic* and mapped *touchy* and *reprimanding*, together with *angry*, onto **A***ngry*[2] as representing different but closely related kinds of negative attitude. For this more balanced 4-class problem AMEN, 1557 words for **A***ngry* (**A**), 1224 words for **M***otherese* (**M**), and 1645 words each for **E***mphatic* (**E**),and for **N***eutral* (**N**), are used, cf. (Steidl et al., 2005). Cases where less than three labellers agreed were omitted as well as those cases where other than these four main classes were labelled. For this AMEN subset, weighted kappa is 0.59.

This sub-sample has been used in several experiments and will be defined as (the basis of) the main 'canonical' samples to be processed, cf. below.

3.4. Chunking into and mapping of labels onto syntactically and emotionally meaningful units

Now we were facing the task of mapping word-based labels onto higher units, first onto turns: a simple 50% threshold – for instance, if an **A** turn has 10 words , then 5 or more

[2] If we refer to the resulting 4-class problem, the initial letter is given boldfaced and recte. Note that now, **A***ngry* can consist, for instance, of two *touchy* and one *reprimanding* label; thus the number of **A***ngry* cases is far higher than the sum of *touchy*, *reprimanding*, and *angry* MV cases.

words have to be labelled as **A** – would be suboptimal because some words, esp. function words, are likely not to be produced in an emotional manner; moreover, a longer turn can consist of one neutral clause, and one emotional clause - then chances are that the whole turn will wrongly be mapped onto neutral.

For the mapping onto turn-based labels, we employed the following strategy: as stop words, fragments and auxiliaries were used; for the turns containing our 6070 AMEN words, this means 17618 words, 3996 turns; stopwords are: 596 fragments, 196 auxiliaries (some words both), i. e. 16856 words remaining.[3] For each turn, we add together the labels given by our 5 labellers (for *n* words, 5 x *n* labels). If the turn is mapped onto neutral, 70% of the labels have to be neutral. (*joyful* and the other spurious labels are not taken into account for this computing.) If 30% or more are non-neutral, then the turn is **A**, **M**, or **E**. If at least 50% of the non-neutral labels are **M**, the turn is mapped onto **M**. If **A** and **E** are equally distributed, the turn is mapped onto **A**. If the turn is neither **A** nor **M**, it is **E**. This simply means that we employ a sort of 'markedness' condition: **M** is more marked than **A**, and **A** is more marked than **E**, and all are more marked than **N**. This yields the following turn-based labels: 868 **A** (21.7 %), 1347 **E** (33.7 %), 495 **M** (12.4 %), and 1280 **N** (32.0 %), summing up to 3990 (100 %) labels = turns.

For the mapping onto 'chunk-based' labels in between word level and turn level, we first annotated the whole database with a coarse syntactic boundary system (main/subordinate clauses, free phrases, dislocations, and vocatives as label especially tuned for these data); we then used similar mapping rules as for the turns. The rules are given explicitly in a structogram in a forthcoming paper.

Our 'turns' are similar to the units used in other studies. As they can consist of up to 53 words, they are not really optimal – we claim that our chunks are. By using different thresholds etc., the chunk size can be adapted to specific needs; the same way, different chunk sizes can be established for finding out how classifiers behave if faced with shorter or longer units. A pivotal characteristic of this solution is that our chunks are syntactically – and by that, semantically – well defined. This is a necessary prerequisite for higher linguistic (deep or shallow) processing in any end-to-end automatic dialogue system.

3.5. Automatic forced alignment per word plus manual correction of this automatic segmentation

We did an automatic forced alignment using the transliteration (i. e., the spoken word chain). Such an alignment is nowadays rather good but of course sometimes erroneous. We therefore decided to have the automatic word segmentation corrected manually for the whole database; the segmentation of the 3990 AMEN turns was cross-checked by the first author.

A corrected reference segmentation allows to exclude wrong segmentation as a source of misclassification, and makes comparisons across approaches more reliable because at least this factor can be kept constant.

3.6. Automatic pitch extraction plus manual correction of this extraction

Historically, pitch has had a prominent position w. r. t. all feature types because of the preponderance of intonation models in the last decades. Even if this might not be mirrored in empirical results, it is of course an important parameter which is, however, notoriously known as impossible to be extracted fully reliably. Databases with corrected pitch values are rare, and we do not know of any other emotion database with such corrected values. We therefore decided to use, in addition to our own pitch detection algorithm (PDA), a well-known frame-based PDA as baseline and correct these values manually. This was done for the 3990 AMEN turns by the first author. More details and differences in classification performance can be found in (Batliner et al., 2007b; Batliner et al., 2007a).

Such manually corrected pitch values do not constitute a ground truth; they are of course biased towards the automatic PDA used. A pitch-synchronous correction was not possible, due to time constraints. Note that even 'objective' measures such as laryngographic recordings are no ground truth: they are close to the signal but not close to perception! However, the corrected values can be used for computing F0 features and be compared to such features based on – sometimes erroneous – automatic PDAs. Again, this helps in keeping constant at least one factor, namely the raw pitch values, and makes comparisons across different approaches of computing pitch features more reliable.

3.7. Further annotations, software, and types of data

In dialogues, the dialogue partner's reaction can be valuable information that can be coded and used in classification. In our scenario, the Aibo does not speak and has no facial gestures. However, we can model its behaviour - whether it is co-operative or not. It seems to be plausible that a non-co-operative behaviour triggers negative reactions to a larger extent than co-operative behaviour. Therefore, we annotated the Aibo's [± co-operative] actions, although, because of the effort needed, only for roughly half of the data. This annotation will not be part of the default distribution but can be made available on a bilateral basis.

In connection with FAU Aibo, two software programs were made available to the community: EDE - Evaluating Decoders using Entropy, and eLabel - Labelling of Emotions.[4] Within the CEICES initiative, a feature encoding scheme has been developed aiming at a full coverage of possible acoustic and linguistic features (low level descriptors and functionals). It is ASCII-based but could easily be converted into some other (e.g., XML) representation. This encoding scheme will be made available on demand.

Apart from the close-talk microphone recordings, there are two more types of recordings: one with the microphone of the video-camera used for protocolling the sessions containing noise and reverberation, and a second one which is

[3] Note that of course, we could find some more stop words, but this would be rather data-driven and not generic so we refrained from that.

[4] http://www5.informatik.uni-erlangen.de/en/our-team/steidl-stefan/free-software-in-humaine/

artificially reverberated. These are not part of the distribution but can be made available on a bilateral basis. (Note that for privacy reasons, the video recordings will not be available.) More details are given in (Schuller et al., 2007).

4. Mandatory benchmarking

Apart from the lack of (freely) available spontaneous, realistic databases in the field of emotion classification – and because of this lack – comparisons of performance across studies, let alone strict evaluations using the very same data such as the ones conducted within the NIST initiative, are practically impossible. Note that even in the case of rather well-defined acted databases such as Emo-DB and DES (well-defined because the classes are given trivially via speaker instructions), researchers often do not select exactly the same cases per class. This makes a comparison of performance across studies impossible. However, in our experience, a convenient selection of sub-samples out from a whole corpus often contributes more to classification performance than the choice of the one or the other feature selection procedure or classifier. Even if this selection is well documented (which is not always the case), it is not clear how much it contributes, if there is no baseline setting. Thus for our database, we want to define and make available in the distribution an 'evaluation setting': a simple two-fold cross-validation which can be computed with rather low effort. As benchmark, we will provide classification results for this constellation which is defined extensionally and thus unambiguous; its processing – in addition to any other processing – will be mandatory.

Of course, the data can and should be exploited in many different ways. Experience tells us, however, that often, data are chosen in a way that results in highest possible recognition rates – and this is what readers remember. By defining obligatory constellations which are, on the same time, simple enough and do not need a high processing effort, we want to establish a sort of benchmark which has to be used by licencees in studies on these data.

5. Concluding Remarks

Using our annotations, the impact of alternative approaches can be pursued (different size of units of analysis, automatically vs. manually extracted values, etc.); seen from a performance point of view, the difference might not be marked; however, adding several small differences might yield larger ones.

The field of automatic classification of emotional user states is still in its infancy, compared to other fields such as automatic speech recognition. This is foremost due to its topic: whereas a word is a word is a word, even in noisy condition, it is neither clear what an emotion is, nor where we can find which emotion in which unit of analysis. All this has to be annotated somehow. This makes it expensive to create databases[5], and databases greater by some order of magnitude would be one prerequisite for standardization in this field. What we hopefully can aim at in the next time to come is thus not a strict evaluations but something like 'islands of standardization': studies dealing, for example, with FAU Aibo can be compared w.r.t. the benchmark constellation.

[5] A (very!) coarse estimate of our expenses for designing, recording, and processing manually FAU Aibo amounts to > 80 K Euros for researchers and students.

6. References

A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann. 2005. Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 489–492, Lisbon.

A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. 2006. Combining Efforts for Improving Automatic Classification of Emotional User States. In *Proceedings of IS-LTC 2006*, pages 240–245, Ljubliana.

A. Batliner, S. Steidl, and E. Nöth. 2007a. Laryngealizations and Emotions: How Many Babushkas? In *Proceedings of the International Workshop on Paralinguistic Speech — between Models and Data (ParaLing'07)*, pages 17–22, Saarbrücken.

A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. 2007b. The Impact of F0 Extraction Errors on the Classification of Prominence and Emotion. In *Proceedings of ICPhS 2007*, pages 2201–2204, Saarbrücken.

A. Batliner, S. Steidl, C. Hacker, and E. Nöth. 2008. Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech. *User Modeling and User-Adapted Interaction*, 18:175–206.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of german emotional speech. In *Proc. 9th Eurospeech - Interspeech 2005*, pages 1517–1520, Lisbon.

Inger S. Engberg, Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard. 1997. Design, recording and verification of a Danish emotional speech database. In *Proc. Eurospeech*, pages 1695–1698, Rhodes.

D. Küstner, R. Tato, T. Kemp, and B. Meffert. 2004. Towards Real Life Applications in Emotion Recognition. In E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems*, pages 25–35, Berlin, Springer.

B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl. 2007. Towards more Reality in the Recognition of Emotional Speech. In *Proc. of ICASSP 2007*, pages 941–944, Honolulu.

S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. 2005. "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency. In *Proc. of ICASSP 2005*, pages 317–320, Philadelphia.

R. Tato, R. Santos, R. Kompe, and J.M. Pardo. 2002. Emotional space Improves Emotion Recognition. In *Proc. ICSLP 2002*, pages 2029–2032.

Emotional Speech Corpus Construction, Annotation and Distribution

Dr. Charlie Cullen, Brian Vaughan, Spyros Kousidis, John McAuley
Digital Media Center, Dublin Institute of Technology, Aungier Street, Dublin 2, Ireland
E-mail: charlie.cullen@dit.ie, brian.vaughan@dit.ie, spyros.kousidis@dit.ie, john@dmc.dit.ie

Abstract

This paper details a process of creating an emotional speech corpus by collecting natural emotional speech assets, analysisng and tagging them (for certain acoustic and linguistic features) and annotating them within an on-line database. The definition of specific metadata for use with an emotional speech corpus is crucial, in that poorly (or inaccurately) annotated assets are of little use in analysis. This problem is compounded by the lack of standardisation for speech corpora, particularly in relation to emotion content. The ISLE Metadata Initiative (IMDI) is the only cohesive attempt at corpus metadata standardisation performed thus far. Although not a comprehensive (or universally adopted) standard, IMDI represents the only current standard for speech corpus metadata available. The adoption of the IMDI standard allows the corpus to be re-used and expanded, in a clear and structured manner, ensuring its re-usability and usefulness as well as addressing issues of data-sparsitiy within the field of emotional speech research.

1. Introduction

Advances in both speech/emotion recognition and emotional speech synthesis largely depend on the availability of annotated, emotional speech corpora. Although it is common that corpora are purpose-built for specific applications or research purposes, it would be desirable to re-use existing corpora. However, there is a lack of widely accepted standards in such areas as audio quality, annotation with metadata in order to perform queries, as well as mutually agreed definitions, as in 'what is emotion?'(Cowie and Cornelius 2003). The work described here is a developing process of emotional asset acquisition, annotation and on-line publishing for emotional rating by end users, which attempts to address some of the above issues, while being flexible in practical issues such as re-usability, standardisation and access. The paper is divided into three parts: (1) A method for obtaining "genuine" emotional speech recordings, namely Mood Induction Procedures (MIP 4) (Gerrards-Hesse, Spies et al. 1994), while recording in a controlled environment; (2) the analysis and annotation of the recorded assets via a purpose-built audio analysis tool (Cullen 2008) and (3) an implementation of the IMDI corpus annotation schema.

2. Genuine Emotional Speech

There are three main forms of asset used in existing speech corpora: simulated assets, broadcast assets and induced assets. A few examples of claimed 'natural' emotional speech databases exist (Scherer and Ceschi 1997;2000; Chung 1999; Douglas-Cowie, Campbell et al. 2003) although the justification for such content is that it is more natural when compared with simulated content. In the majority of cases what is termed 'natural' emotional assets are obtained from a broadcast source (mainly television) (Douglas-Cowie, Campbell et al. 2003). However it can be argued that assets obtained from such sources may not be natural or contain genuine emotional content.

2.1 Simulated Assets

Corpora consisting of simulated assets use acted emotional states, read texts and imagined/recalled emotional situations (Banse and Scherer 1996; Enberg 1997; Amir 2000; Kienast 2000; Pereira 2000). However very little is actually known about how simulated emotion compares to natural emotion (Douglas-Cowie, Campbell et al. 2003). Simulated emotion that involves reading from a text is not a spontaneous expression of emotion with read speech having distinct characteristics from spontaneous speech (Johns-Lewis 1986), with vowel substitution and reduction being more likely to occur in spoken as opposed to read speech (Van Bael 2004). Emotional states can be considered to be an important factor in maintaining and negotiating social interaction and relationships (Cornelius 2000), communicating information about our intentions and possible behaviour to those around us: they compel us to action and regulate social communication (Plutchik 2001). Simulated assets are often non-interactive (Banse and Schrerer 1996; Enberg 1997; Amir 2000; Kienast 2000; Pereira 2000), consisting of monologues with little or no interaction from other agents. The neglect of the social dimension of emotional speech means that obtained assets may contain only a limited range of emotions.

Numerous commentators have argued that there are fundamental biological and physiological aspects to emotion (Darwin 1872; James 1884; Bindra 1969; Frijda 1988; McGuire 1993) with Johnstone (Johnstone 1996) arguing that emotion can induce changes in speech that the speaker cannot control and that these changes reflect the underlying physiological changes taking place. It is debatable whether these uncontrollable changes are present in simulated emotional speech, therefore,

simulated emotions may well be nothing more than a resemblance of real emotional states (Pugmire 1994). Thus the voluntary and non-spontaneous nature of simulated emotion may undermine its authenticity and its suitability as a method of obtaining natural emotional speech assets.

2.2 Broadcast Assets

Some corpora use assets obtained from broadcast sources, mainly television (Chung 1999; Douglas-Cowie, Campbell et al. 2003), the justification being that they are 'natural' compared to simulated assets (Douglas-Cowie, Campbell et al. 2003). Some of the problems associated with the use of simulated assets are also of concern in using broadcast assets. Furthermore, it can be argued that any broadcast is a performance, as the speakers are usually very aware of the recording process taking place. It is recognised in anthropological research that the presence of a researcher and equipment may cause people to act differently or even feel constrained in what can be said and done (Geer 1957; Gottdiener 1979). It is possible that this distortion and constraint means that televised emotional displays, like simulated emotion, may only be a facsimile of real emotion. The only way to prevent this distortion is to conceal the equipment and covertly record subjects; however this is a highly questionable practice and ethically unsound. The distorting effect may lessen over time as subjects become used to being recorded (Erickson 1982). This would suggest that it would be more relevant to use clips taken from the middle or towards the end of a televised program as opposed to clips taken from the start. However, there is an inherent perceptual bias to the recording process (Bellman 1977). This perceptual bias is inherent in the subjective decisions of the cameraman, the director, the producers and the editor and it cannot be known how this affects the final outcome of a broadcast piece.

2.3 Audio Quality

Assets taken from broadcast sources can be of varying audio quality, as 'broadcast quality' is a term rather than a definition; one cannot assume that assets obtained from broadcast sources are of uniform quality. Audio quality will also vary depending on the nature of the program, whether it is recorded in a studio or outside in public spaces (as many reality television programs are). Various other factors will affect the audio quality: noise from studio audiences, people talking across each other and environmental noise from outside broadcasts. The equipment used will also affect the sound quality: different broadcast situations may use different recording apparatus (microphones, cameras etc) and methods. The greatest single advantage of simulated assets is the potential for control of the recording environment, such that most simulated assets are obtained using studio equipment and conditions. The huge variation in recording quality found in other types of corpora (such as those using broadcast assets) precludes the definition of cohesive standards, and thus simulated assets are often preferred for this reason.

2.4 Natural Assets and Mood Induction Procedures

In order for assets to be considered natural for the purpose of analysis, the authors argue that they should be derived from non-simulated and non-broadcast sources with audio quality being of paramount importance. The induction of natural emotional responses in a laboratory environment, thus ensuring audio quality can be maintained, is achieved through the use of Mood Induction Procedures. Mood Induction Procedures (MIPs) are procedures that are designed to induce specific emotional states in a test subject within a controlled situation. The Success/Failure MIP (Forgras 1990; Henkel 2004) uses false feedback (positive or negative) concerning a subject's performance in a test that they believe is testing their cognitive ability. By placing subjects in a situation where certain needs are activated, such as the need to succeed at a certain task, frustrating or aiding the subject in the attainment of their need can induce emotional states. While other MIPs have been found to be more successful in some cases (Gerrards-Hesse, Spies et al. 1994), their effectiveness may be overestimated due to demand effects (Westermann 1996). Demand effects pose a problem to the validity of MIPs due to the fact that participants may guess the purpose of the procedure (to elicit emotional responses) and so pretend to be experiencing the desired emotion. Any instruction given regarding required emotional states can cause a demand effect. The Success/Failure MIP avoids the creation of demand effects: the true nature of the experiment is not evident and can be further disguised if needed. Participants are engaged in a task and can be led to believe that the completion of the task is the purpose of the experiment. The use of false feedback, either positive or negative, further conceals the true purpose of the experiment. The use of a task based Success/Failure MIP may remove the subjective nature associated with some other MIPs, and allows the researcher to control and manipulate the experiment in greater detail. By frustrating or aiding the subjects in their task, without their knowledge, they can be guided towards natural negative or positive emotional states without being aware that a certain emotional state is required, thus avoiding the creation of demand effects.

2.4.1. MIP Audio Quality

The use of Mood Induction Procedures to stimulate emotion has the potential for the same recording conditions to be applied as with simulated assets. The difficulties associated with such conditions using MIPs are related to the concealment of recording equipment to avoid revealing the true purpose of the experiment prior to commencement. In Kehrein's experiment (Kehrein 2002), the fact that the participants were seated in separate sound proofed rooms, allowed the conversational interaction to be recorded as two separate high quality audio channels. This allowed both sides of the conversation to be analysed, including overlaps. Participants were aware of the presence of the recording equipment but believed it was used for them to communicate with each other.

A task-based MIP offers a high degree of control, either hindering or aiding participants, while the use of separate sound proofed rooms enables high quality audio assets to

be obtained. This approach ensures that obtained assets are natural, compared to simulated and broadcast assets, while the co-operative nature ensures the social aspect of emotional expression is not neglected. The resulting emotional assets can be claimed to be natural and spontaneous, arising out of the manipulation of the task and the interaction of the participants as opposed to voluntary or knowingly coerced attempts to generate emotional states.

2.5 MIP Experiment

Taking into consideration the arguments presented above, an MIP was devised using modern games consoles and games, in conjunction with sound isolation booths (Vaughan 2007).The main advantage of these console systems is that a large amount of the games are usually designed with extensive multiplayer options that are cooperative and/or competitive in nature. Computer games have been used before by Jonhnstone (Johnstone, Reekum et al. 2005) and as far back as 1978 by Isen et. Al (Isen 1978) to elicit emotional responses. Johnstone in particular noted that they are particularly suited to this task due to the fact that they can easily be changed and manipulated in order to suit the experiment. Little or no external manipulation is necessary as modern game design focuses very much on immersing the gamer in the gaming world while the majority of games have been designed to be competitive and challenging, usually with an emphasis on competitive goal achievement. The overall game play and style of most games is therefore conducive to inducing emotional states in participants. External manipulation can be achieved, where needed, through unplugging a participant's game controller, changing the time limit, giving false information regarding the amount of time left or through offering a cash or material reward (for elated states).

Participants in the experiment are aware they are being recorded, all must sign a consent form giving permission for the recordings to be used for research purposes. However the true nature of the experiment is not revealed; participants are led to believe that competitive gaming or in game communication is being studied, thus minimizing the chance of a demand effect manifesting. The audio is recorded at 24bit/9Khz and once enough recordings have been attained, it is then analysed and annotated.

3. Analysis and Annotation

The audio recordings obtained from the gaming MIP are segmented into short phrases/clips. At present this is done by hand using a digital audio editor. However it is envisaged, and work is being undertaken in this area, that this will eventually be a semi-automatic or automatic process. These audio clips are then processed using the LinguaTag application in preparation for inclusion in the speech corpus.

3.1 LinguaTag

LinguaTag (Cullen 2008) is a purpose-built application, has been developed for the acoustic analysis and first stage of annotation (tagging) of the corpus. LinguaTag is written in Eiffel (Software 2008) and makes use of the PRAAT (Boersma and Weenink 2006) engine to obtain low-level acoustic data from the recorded signal, while providing a user-friendly interface for transcription, segmentation, labeling and emotional rating of the sound clips . The low-level analysis includes automatic vowel identification, with pitch, intensity and formant contours, as well as voice quality measures (Johnstone and Scherer 1999; Gobl, Bennett et al. 2002) calculated for each vowel. Separate tiers in the annotation schema allow for acoustic analyses of larger clips, as well as the other linguistic annotations mentioned above. In addition, three sliders are available (pitch, intensity, duration) for setting thresholds for detection of stressed vowels. Emotional rating is performed using a circumplex model, comprised of two axes (activation and evaluation), adapted from Scherer (Scherer 1984; Vaughan 2007). LinguaTag outputs this data in a separate XML file, following the SMIL format (Consortium 2008; XML 2008). This XML file is then uploaded with the original WAV file and an MP3 file for use in future online listening tests.

3.2 IMDI Corpora

Consideration has been given to the annotation schema itself, as the existence of metadata is arguably as crucial as the content of the corpus: metadata can be used to query data in a corpus, thus expanding its usability and re-usability. Developing powerful emotional speech technology applications and in-depth analysis in emotional speech research require ever larger amounts of data, both to overcome problems such as data-sparsity (Xiao, Dellandrea et al. 2005) and to enable use of the most appropriate data available. Therefore, corpora need to be sufficiently annotated for such queries to be possible, and there has to be a standardisation of the annotation form, to allow for easy universal access.

Unfortunately, there is a lack of standardisation for annotating speech corpora, particularly in relation to emotive content. The only cohesive attempt at corpus metadata standardization performed thus far has been by the EAGLE/ISLE consortium (ISLE 2003), which has led to the development of the ISLE Metadata Initiative (IMDI). Although not a comprehensive (or universally adopted) standard, IMDI represents the only current standard for speech corpus metadata available. For this reason, it was decided to implement the IMDI standard within the speech corpus detailed in this paper in order to maintain as cohesive a standard as possible within current developments. The IMDI schema is extensive and so it was decided that initially only the four higher tiers of the schema (Project, Session, Actor and Content) would be implemented. It was felt that these were the most relevant elements of the schema to the corpus. This does not preclude the inclusion of other elements of the schema from being implemented at a later date should it be deemed necessary.

3.3.1. Implementation

The implementation of the IMDI annotation schema is structured as follows: A project groups together different bundles of sessions. A session is defined as the common bundle for linguistic events within IMDI metadata, and thus all speech assets are defined relative to a specific session. This allows an audio clip to be taken from a

longer recording for specific analysis, while still retaining the same overall metadata as all other files in that session bundle. Within each session, actors, i.e. participants in the recordings, are documented (with anonymity preserved at all times for ethical reasons) so that database queries involving geographical information or age can be performed. The content metadata relates to specific activities for a given session, such as the type of emotional content (induced, acted, etc) or other types of content categorization (the vocabularies are open for some of the tags). Finally, the asset metadata relates to the low-level acoustic information and the linguistic and emotional annotation that is performed by LinguaTag in SMIL format.

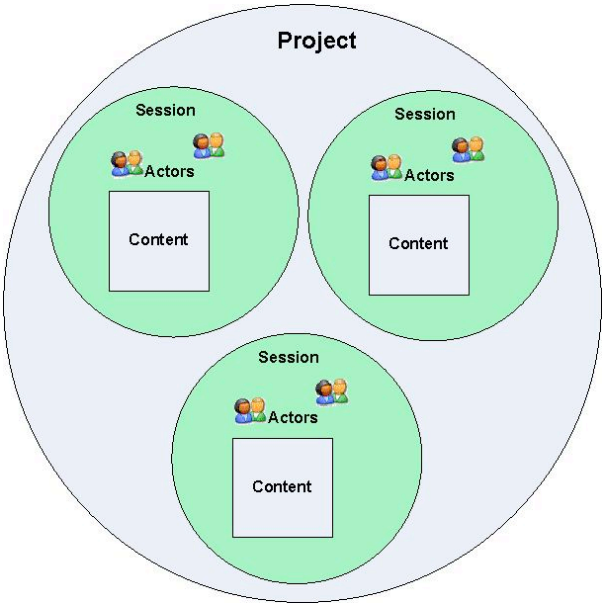


Figure 1: Example block diagram of the IMDI schema organisation. In this example, 3 separate session bundles are grouped logically under a single project.

This approach is advantageous in that the definition of a particular project allows various sessions to be grouped in a logical form. Thus, in the case of the emotional speech corpus described in this paper, all sessions are organised relative to the project. Grouping sessions logically, allows for future expansion of the speech database to include other corpora developed for different purposes. The session definition provides a convenient way to group assets for analysis, allowing assets taken from different experiments to be assessed either in isolation or within a wider common context. The definition of an actor(s) within a session is a very useful aspect of the IMDI standard, as it allows the various participants in a speech recording to be documented for later consideration. In many instances, actor details may be vague and non-specific to ensure that ethical standards are adhered to (this is given as an option for each testing participant). However, as mentioned more detailed actor information would be of use for certain types of queries. Future work may consider the multi-lingual definition of assets within a corpus for analysis, and so actor information would be crucial for this.

The annotation schema provides the flexibility of querying assets for different properties, such as speaker characteristics, emotional dimensions (Cullen 2006; Vaughan 2007) (e.g. 'only negative' or 'extremely active'), or certain audio quality. In addition, the overhead associated with tagging the audio files is greatly reduced by the use of automation functionality (auto-suggest) during the tagging process. Metadata previously entered can be reused, e.g. metadata that is the same for the whole session need only be entered once. Similarly, any metadata shared between any number of assets in any combination, need not be re-entered, as it is available through the autosuggest functionality.

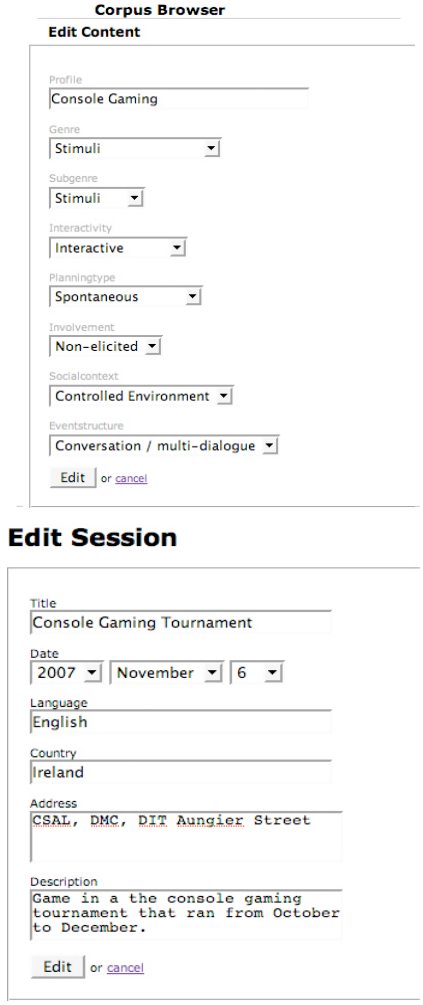


Figure 2: Screenshots of the Session and Content screens

There were, from the outset, several considerations that helped to define the technical architecture of the corpus. Firstly, the prototype must provide editors with the ability to insert assets, in the form of WAV files, and related Linguatag data, in the form of SMIL files. The prototype must parse the SMIL file and populate the corresponding database tables. The corpus, therefore, necessitates a storage layer or database as a persistent back-end. Secondly, editors require remote access to corpus assets. This allows for the addition, deletion and alteration of corpus assets and related metadata. At first, each asset was to be uploaded and annotated individually. However, following initial trials, it was decided to provide the ability for batch uploads, thereby allowing an editor to upload several assets at the any one time. In this case, each asset is annotated with the same metadata.

4. Conclusion

This paper considered a method for obtaining natural emotional assets and annotating them as part of a speech corpus. MIPs were determined as the best method for obtaining natural emotional speech assets. A gaming based MIP experiment was developed to elicit natural emotional responses from participants. In order to analyse these assets an application, LinguaTag, was developed (Cullen 2008), providing a SMIL file with detailed acoustic information that can be parsed by a relational database. The IMDI corpus standardisation was adopted and implemented in order to annotate the assets and provide a clear and concise method by which the data could be structured in a 3-tiered system. This approach goes someway to avoiding data-sparsity and improving the inter-operability of the corpus.

At time of writing, the corpus contains over 150 fully annotated and tagged assets, and this figure is intended to grow. There is no defined headroom for the size of the corpus, but the experimental criteria, recording conditions and annotation metadata will be upheld in all future work. An on-line listening tool is also being developed and tested and will be the method by which on-line listening tests are carried out to rate the emotional dimensions of the assets. The intention of the online rating system is to obtain a statistical definition of emotional dimensions for each clip in the corpus, and rate each clip both in terms of its dimensional values and also the confidence rating for that clip. Thus, a clip which has been rated by more listeners will be defined as having a higher confidence level relative to its emotional dimension values, allowing statistical analysis to be performed on groups of assets in as robust a manner as possible.

5. Acknowledgements

This work was funded by the SALERO project. Special thanks to Charlie Pritchard and Evin McCarthy.

6. References

Amir, N., Ron, S., Laor, N. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. ISCA ITRW on Speech and Emotion, Belfast.

Banse, R. and K. R. Scherer (1996). "Acoustic profiles in vocal emotion expression." Journal of Personality and Social Psychology **70**(3): 614-636.

Banse, R. and K. R. Schrerer (1996). "Acoustic profiles in vocal emotion expression." Journal of Personality and Social Psychology **70**(3): 614-636.

Bellman, B. L., & Bennetta Jules-Rosette. (1977). A Paradigm for looking. Norwood, Ablex Publishing.

Bindra, D. (1969). "A unified interpretation of emotion and motivation." Annals of the New York Academy of Science. **159**: 1071-1083.

Boersma, P. and D. Weenink (2006). Praat: doing phonetics by computer.

Chung, S. (1999). Vocal expression and perception of emotion in Korean. 14th International Conference of Phonetic Sciences, San Fransisco, USA.

Consortium, W. W. W. (2008). "Synchronized Multimedia." from http://www.w3.org/AudioVideo/.

Cornelius, R. R. (2000). "Theoretical approaches to emotion." Speech Emotion **1**: 3-10.

Cowie, R. and R. R. Cornelius (2003). "Describing the emotional states that are expressed in speech." Speech Communication Special Issue on Speech and Emotion **40**(1-2): 5-32.

Cullen, C., Vaughan, B. ,Kousidis, S., Wang, Yi ., McDonnell, C. and Campbell, D. (2006). Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction International Conference on Multidisciplinary Information Sciences and Technologies Extremadura, Merida.

Cullen, C., Vaughan, B., Kosidis, S. (2008). LinguaTag: an emotional speech analysis application. Accepted paper at: The 12th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2008. Orlando, Florida, USA.

Cullen, C., Vaughan, B., Spyros,. K. (2008). LinguaTag: an emotional speech analysis application. 12th World Multiconference on Systemics, Cybernetics and Informatics (WMSCI 2008). Orlando, Florida. USA.**:** 7.

Darwin, C. R. (1872). The Expression of the Emotions in Man and Animals. London., Albermarle.

Douglas-Cowie, E., N. Campbell, et al. (2003). "Emotional speech: towards a new generation of databases." Speech Communication Special Issue Speech and Emotion **40**(1-2): 33–60.

Enberg, I. S., Hansen, A.V., Anderson, O., Dalsgaard, P. (1997). Design, recording and verification of a Danish Emotional Speech Database. Eurospeech '97, Rhodes, Greece.

Erickson, F., and Schultz, J. (1982). The Counsellor as Gatekeeper: Social Interaction in Interviews. Language, Thought and Culture: Advances in the Study of Cognition. E. Hammel. New York, Academic Press.

Forgras, J. P. (1990). "Affective influences on individual and group judgements ." European Journal of Social Psychology **20**: 441-453.

Frijda, N. H. (1988). "The Laws of Emotio." American Psychologist **43**(5).

Geer, B. a. H. S. B. (1957). "Participant Observation and Interviewing: A Comparison." Human Organization **16**(3): 28-32.

Gerrards-Hesse, A., K. Spies, et al. (1994). "Experimental inductions of emotional states and their effectiveness: A review." British Journal of Psychology **85**: 55-78.

Gerrards-Hesse, A., K. Spies, et al. (1994). "Experimental inductions of emotional states and their effectiveness: A review." British Journal of Psychology **85**(1): 55-78.

Gobl, C., E. Bennett, et al. (2002). Expressive Synthesis: How Crucial is Voice Quality? IEEE Workshop on Speech Synthesis, Santa Monica, CA (USA).

Gottdiener, M. (1979). "Field Research and Video Tape."

Sociological Inquiry **4**(49): 59-66.
Henkel, M., J., Hinsz, V. (2004). "Success and failure in goal attainment as a mood induction procedure." Social Behavior and Personality **32**(8): 715-722.
Isen, A., Shalker, T., Clark, M., Karp., L. (1978). "Affect, acessibility of Material in Memory, and Behavior: A Cognitive Loop?" Journal of Personality and Social Psychology **36**(1): 1-12.
ISLE. (2003). "IMDI (ISLE Metadata Initiative), Metadata Elements for Session Descriptions." Draft Proposal Version 3.0.3. from http://www.mpi.nl/IMDI/Schema/IMDI.
James, W. (1884). "What is an emotion?" Mind **9**: 188-205.
Johns-Lewis, C. (1986). Prosodic differrentiation of discourse modes. Intonation in Discourse. C. Johns-Lewis. San-Diego, College Hill Press**:** 199-220.
Johnstone, T. (1996). Emotional Speech Elicited using computer games. Spoken Language, ICSLP 96. Proceedings., Fourth International Conference on, Philadelphia, PA, USA.
Johnstone, T., C. M. v. Reekum, et al. (2005). "Affective speech elicited with a computer game." Emotion(5): 513-518.
Johnstone, T. and K. R. Scherer (1999). The Effects of Emotions on Voice Quality. XIV Int. Congress of Phonetic Sciences, San Francisco.
Kehrein, R. (2002). The prosody of authentic emotions. Speech Prosody, Aix-en-Provence, France.
Kienast, M., Sendlmeier, W.F., (2000). Acoustical analysis of spectral and temporal changes in emotional speech. ISCA ITRW on Speech and Emotion, Newcastle, Belfast, Textflow.
McGuire, T. T. (1993). Emotion and behaviour genetics in vertebrates and invertabrates Handbook of Emotions. M. Lewis, Haviland, J.M. New York, Guilford Press.
Pereira, C. (2000). Dimensions of emotional meaning in speech. ISCA ITRW on Speech and Emotion, Newcastle, Belfast, Textflow.
Plutchik, R. (2001). "The Nature of Emotions." American Scientist **89**(4): 344-350.
Pugmire, D. (1994). "Real Emotion." Philosophy and Phenomenological research **54**(1): 105-122.
Scherer and Ceschi (1997;2000). "Geneva Airport Lost Luggage Study." Motivation and Emotion **21**: 211-235.
Scherer, K. R. (1984). On the nature and function of emotion: A component process
approach. Approaches to emotion. K. R. Scherer and P. Ekman. Hillsdale, NJ, Erlbaum**:** 293-317.
Software, E. (2008). "Eiffel Software Home Page." Retrieved February, 2008.
Van Bael, C., van den Heuvel, H., Strik, H. (2004). Investigating Speech Style Specific Pronunciation Variation in Large Spoken Language Corpora Large Spoken Language Corpora. Proceedings of Interspeech (ICSLP) Jeju, Korea.
Vaughan, B., Kosidis, S., Cullen, C., Wang, Yi. (2007). Task-Based Mood Induction Procedures for the Elicitation of Natural Emotional Responses. The 4th International Conference on Cybernetics and Information Technologies, Systems andApplications: CITSA 2007 Orlando, Florida.
Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). "Relative effectiveness and validity of mood induction procedures: a meta analysis." European Journal of Social Psychology **26**: 557-580.
Xiao, Z., E. Dellandrea, et al. (2005). Features extraction and selection for emotional speech classification. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), Como, Italy.
XML, W. W. W. C. (2008). "Extensible Markup Language." from http://www.w3.org/XML/.

Emotions in a Corpus of Human and Computer Keyboard-to-Keyboard Tutoring Sessions

Farhana Shah[1], **Martha Evens**[2]

[1]Institute of Information Technology, Quaid-i-Azam University, Islamabad, Pakistan

[2]Department of Computer Science, Illinois Institute of Technology, 10 West 31st Street, Chicago, IL 60616

E-mail: farhanasha@yahoo.com, evens@iit.edu

Abstract

We have developed a corpus of 75 human tutoring sessions and 139 machine tutoring sessions. The human tutors are Joel Michael and Allen Rovick, Professors of Physiology at Rush Medical College, tutoring students about the baroreceptor reflex system, the negative reflex system that controls blood pressure in the human body. All sessions were carried out keyboard-to-keyboard fashion with tutors and students in two separate rooms. The machine tutor, CIRCSIM-Tutor, was designed to carry out a natural language dialogue with the student; its strategies, tactics, and language are modelled on the human sessions. The students in both human and machine sessions were first year medical students at Rush Medical College. Our goal is to determine how to make CIRCSIM-Tutor detect student emotion and respond to it. We have marked up expressions of emotion in the human sessions extensively. Contrary to the Media Equation of Reeves and Nass, the students express much more emotion in the human sessions than in the machine sessions and the emotions expressed are much more positive in those sessions. As anyone would expect, the human sessions show much more wit, charm, and flexibility. The differences in expressed emotion may disappear as the language abilities of CIRCSIM-Tutor improve.

1. Introduction

Almost twenty years ago we set out to build an Intelligent Tutoring System called CIRCSIM-Tutor that could carry on a natural language dialogue with students. This plan was first conceived by Joel Michael and Allen Rovick, Professors of Physiology at Rush Medical College, who had already authored several CAI systems for their students. Medical students, several years away from their required one semester of college-level physics, often have difficulty understanding negative reflex systems, so we chose to focus on the baroreceptor reflex, which controls blood pressure in the human body. The two domain experts had already spent a lot of time tutoring students on this material and also running small group problem-solving sessions. Our goal was to model the kind of tutorial dialogue that the experts carried on with their students, so our first step thought was to make audio-tapes of several spoken tutoring sessions. The problems involved in the process of transcribing these sessions and the questions raised about the representation of the smiles, frowns, raised eyebrows, laughter, and groans that accompanied the dialogue convinced the experts to abandon spoken dialogue in favor of keyboard-to-keyboard dialogue with student and tutor in separate rooms. They decided that this mode of keyboard tutoring would force them, and, of course, their students, to express all the interaction in words and thus provide us with much more appropriate examples of language to model in our system. This experiment was eventually successful; students do learn from the system (Evens & Michael, 2006; Michael et al., 2003).

2. Dialogue Transcripts

Over the years Joel Michael and Allen Rovick carried out 75 sessions, most of them approximately one hour in length; the students in all cases were first-year students at Rush Medical College. All of these transcripts were captured using a version the Computer Dialogue System (CDS) developed by Jun Li (Li et al., 1992). CDS controls turn-taking, so that the tutor and the student cannot both type at once. Each party to the dialogue must explicitly type a contribution, then yield the turn. A record of the current portion of the dialogue appears on both screens; the complete transcript is preserved as a file on the tutor's hard-drive. Either party may interrupt the other, but only by asking for permission and receiving it. Once the session is over a numbering program goes through and labels each sentence with an indication of the person speaking, the number of the turn within the session, and the number of the sentence within the turn, as shown in Example 1.

Example 1:

K51-tu-44-1: OK, and what determines the intracellular concentration of Ca?
K51-st-45-1: The reflex.
K51-st-45-2: Oops.
K51-tu-46-1: so, what do you want to predict for IS?
K51-st-47-1: IS 0
K51-tu-48-1: right.

After the input understanding module was replaced by a new one using a semantic grammar approach (Glass, 1999; Glass & Evens, 2008), the experts decided that the system was good enough to do a large scale trial of CIRCSIM-Tutor with medical students. We wound up

with 38 transcripts of computer tutoring from November, 1998, and 35 from November, 1999. The system was then made routinely available in a student laboratory. In Fall, 2002, we were able to collect 66 more transcripts of students interacting with a somewhat improved version.

3. Transcript Markup

During the analysis phase we marked up the human keyboard sessions in SGML style in a number of different ways, looking for tutoring strategies and hints, and at the syntax used by the tutor and the student. The markup that may possibly be of interest to the participants of this workshop involves student initiatives (Shah et al., 2002), student hedges, and student affect (Bhatt et al., 2004). A detailed discussion of the whole range of markup can be found in (Kim et al., 2006).

Since the tutor begins with an agenda and gives the student a problem to solve, the tutor has the initiative most of the time. Shah set out to study those occasions when the student does not try to answer the tutor's question, but, instead, takes the initiative by asking a question, proposing a mini-theory and asking for confirmation, challenging something the tutor said, or asking for conversational repair. She marked up student initiatives and tutor responses as shown in Example 2.

Example 2:
K12-tu-87-2: How does the reflex manage to lower TPR?
K12-st-88-1: Dilation of blood vessels
K12-tu-89-1: And how does it accomplish that?
<S-Init, Goal = RFI, Form = Fragment, Focus =Causal Reasoning, Hedged=Y>
<S-HEDGES-SENTENCE,TYPE=EITHEROR, FUN = INIT>
K12-st-90-1: Either decreased symp.
K12-st-90-1: Or increased para.
K12-st-90-2: (did i reverse it)
</S-HEDGES-SENTENCE>
</S-Init>
<T-Resp, Goal = EXP, Form=Declarative, Delivery Mode = Monologue>
K12-tu-91-1: There's practically no parasympathetic innervation of …
</T-Resp>

4. Marking up Emotions

Bhatt looked at the student dialogues with the computer tutor and decided that it was imperative to change the way that CIRCSIM-Tutor responds to student expressions of emotion as illustrated in Example 3.

Example 3:
CIRCSIM-Tutor: What are the determinants of SV?
Student: I hate computers.
CIRCSIM-Tutor: Please respond with prediction table parameters.

He decided to mark up the student affect in the 25 most recent human tutoring sessions: K52-K76 and look at how the human tutors responded. (Example 4 is Example 1 with affect markup added.)

Example 4:
K51-tu-44-1: OK, and what determines the intracellular concentration of Ca?
K51-st-45-1: The reflex.
<S-SHOWS-AFFECT, TYPE=APOLOGY, FUN=ANS>
K51-st-45-2: Oops.
</S-SHOWS-AFFECT>
K51-tu-46-1: so, what do you want to predict for IS?
K51-st-47-1: IS 0
K51-tu-48-1: right.

The markup indicates that the student is expressing an emotion (apologizing for an answer that is clearly not what the tutor wanted, though correct) and that this input functions as part of an answer rather than a student initiative. Unfortunately, Bhatt's emotion typology is entirely ad hoc (see Table 1).

Affect Type	Frequency	Examples
Contemplation	19	Hmmm. Well. I am thinking. Um.
Apology	18	Sorry. Talk about a stupid, careless error. Oops!
Gratitude	14	Thank you.
Realization	14	Aah. Oh.
Comprehension	12	I get it. Aha.
Confusion	10	I'm a bit confused. I'm having difficulty. Got mixed up.
Feedback	6	This has been helpful. That makes sense.
Curiosity	2	I'm curious. I wonder ….
Courtesy	1	Good morning. Good-bye.
Amazement	1	Wow.
Amusement	1	Ha. Ha.
Pain	1	Ouch.
Frustration	0	Let's proceed. Go on. I will ask Dr. Michael.

Table 1: Bhatt's Emotion Types Ordered by Decreasing Frequency in 25 Expert Human Sessions (K52-K76).

Bhatt found that students used many fewer expressions of affect with CIRCSIM-Tutor than with human tutors, and that those that they do use are almost all negative, while those that they used in human sessions were much more positive (Bhatt et al., 2004).

We provide some examples of markup of the more common categories of affect. The examples that Bhatt categorized as contemplation come in two flavors; one kind, shown in Example 5, has a hesitation marker of some kind; the other, shown in Example 6, includes "well," which suggests that the student does not expect the tutor to like this input. Both kinds seem to inspire the tutor to give some kind of explanation.

Example 5:
<S-SHOWS-AFFECT, TYPE=CONTEMPLATION>
K72-st-42-1: hmmm...by not taking up as much Ca into the sarcoplaxmic reticulum
</S-SHOWS-AFFECT>

Example 6:
<S-SHOWS-AFFECT, TYPE=CONTEMPLATION>
K72-st-60-1: well, we wanted to know would happen immediately, not after subsequent steps.
</S-SHOWS-AFFECT>

The categories GRATITUDE and REALIZATION are tied for third place in terms of frequency in the 25 sessions counted here. Recognizing GRATITUDE is relatively easy; some form of the word "thank" appears every time, as in Example 7.

Example 7:
<S-SHOWS-AFFECT, TYPE=GRATITUDE>
K75-st-95-1: Thanks
</S-SHOWS-AFFECT>

Realization, shown in Example 8, appears in a number of different guises. It seems most important to recognize and respond to the ones where the student needs help.

Example 8:
<S-SHOWS-AFFECT, TYPE=REALIZATION>
K64-st-18-2: wait – I don't understand
</S-SHOWS-AFFECT>

The category that Bhatt labeled comprehension is often combined with other types of affect, as in Example 9.

Example 9:
<S-SHOWS-AFFECT, TYPE=COMPREHENSION>
K67-st-56-1: Thanks, I think I understand
</S-SHOWS-AFFECT>

We classify as CONFUSION examples where the student uses the words *difficult*, *confuse*, *confusion*, *trouble*, or *hard*. The student in Example 10 makes it easy by using two of these signals.

Example 10:
<S-SHOWS-AFFECT, TYPE=CONFUSION>
K67-st-50-1: Yes- I seem to have trouble relating TPR and MAp(I get them confused).
</S-SHOWS-AFFECT>

Students realize that people want feedback (Example 11). The computer tutor never sees it.

Example 11:
<S-SHOWS-AFFECT, TYPE=FEEDBACK >
K52-st-54-2: This was fun, I really like this interactive style of learning.
</S-SHOWS-AFFECT>

The students almost never express frustration with the expert tutors, but we included this category in our table because we do see it with novice tutors, as in Example 12. Students working with novice tutors, who are more advanced medical students, challenge those tutors often, and even, as in this case, reject their explanations.

Example 12:
<S-SHOWS-AFFECT, TYPE=FRUSTRATION>
N2-st-65-1: I will refer that question to Dr. Michael.
</S-SHOWS-AFFECT>

5. Emotions in CIRCSIM-Tutor Sessions

Student expressions of emotions in the CIRCSIM-Tutor sessions are very different from those we see in the human tutoring sessions. We probably should not have been surprised. Bozena Thompson (1980), in one of the first studies of human responses to natural language communications with a computer, found that the navy personnel using her system cursed it out regularly, although they told her that the system was useful and effective. This is consistent with the results of a Wizard-of-Oz experiment reported by Barbara Fox (1993). Fox found that the students were pleasant and polite to human tutors, but were extremely rude to the same tutors when they interacted with them over a computer link and were told that they were communicating with a computer.

Emotions in the 1998 Sessions

In November, 1998, 50 first-year students at Rush Medical College used CIRCSIM-Tutor for an hour in a morning laboratory; 24 students worked in pairs so we wound up with 38 session transcripts containing a total of 1801 student turns. Several students typed *help* and received a hint as in Example 13 below.

Example 13:
T: What is the other determinant of Stroke Volume?
S: help
T: Please respond with prediction table parameters
S: cvp

The student can obtain the same kind of result from typing a question mark as in Example 14.

Example 14:
T: What parameter in the prediction table represents preload?
S: ?
T: Please respond with prediction table parameters
S: cvp

So it is not clear that *help* should be classified as expressing emotion. There are also places where the student is almost certainly testing the system, as in Example 15.

Example 15:
T: By what mechanism is Total Peripheral Resistance controlled?
S: by local control or myogenic mechanism l
T: Is the mechanism of control neural or physical?
S: dent
T: Is the mechanism of control neural or physical?
S: both
T: Is the mechanism of control neural or physical?
S: physical

Sometimes the student seems definitely angry as in Example 16. Fortunately, the system recognizes "+" as meaning increased.

Example 16:
T: What is the correct value of Cardiac Output?
S: eat shit
E: It should be increased, decreased, or unchanged
S: kiss my ass
E: It should be increased, decreased, or unchanged
S: +

5.2 Emotions in the 1999 Sessions

In November, 1999, we tried a new version of CIRCSIM-Tutor out with 42 first-year medical students; 14 used the program in pairs, so we wound up with 35 separate transcripts. We made a number of changes to the parser from the year before: we improved the spelling correction algorithm and we recognized a number of correct but previously unexpected inputs that the 1998 version of the program erroneously stigmatized as wrong. We would like to believe that the angry student input disappeared because we had succeeded in making the system less frustrating to use; it seems more likely that this was the result of running a big experiment with a new system; we had faculty and Computer Science students patrolling the room trying avert system crashes or get the system back up if it crashed (it didn't). We found only one definite expression of emotion (see below in Example 17), but a number of places that may be deliberate nonsense inputs or may be just wrong answers. We list the suspicious ones in Table 2.

Example 17:
T: What is the other determinant of Mean Arterial Pressure?
S: reflex
T: Please respond with prediction table parameters
S: -
T: Please respond with prediction table parameters
S: clueless
T: Please respond with prediction table parameters
S: heart rate

It is clear that a human tutor would have responded to that sad "clueless." Before the 2002 experiments we changed the system to stop the process and give the student the answer, if the student receives two responses in a row describing what input the system is expecting and still does not respond with appropriate input. This solution avoids several of the other problems that appeared in the 1999 transcripts.

Sess	CIRCSIM-Tutor Question	Student Ans.
8	What are the determinants of Stroke Volume?	kl
31	What is the other determinant of Stroke Volume?	io
31	Which determinant is dominant in this case?	-
16	Which determinant is dominant in this case?	jj
19	What stage must the value of Cardiac Output follow in SS?	0
19	What is the correct value of Cardiac Output?	t>

Table 2: More Possibly Affective Responses to CST in November, 1999

5.3 Emotions in the 2002 Sessions

In November, 2002, most of the first year class used CIRCSIM-Tutor for one hour in a regularly scheduled laboratory session. Some students worked alone, some in pairs, so we wound up with only 66 transcripts. Interpreting affect very broadly, we found 20 instances of affect in these 66 sessions, far fewer than the 88 instances of affect in the 25 human tutoring sessions that are counted in Table 3. Even more striking, only 12 sessions (18%) contained any affect at all, as opposed to 22 (88%) of the human tutoring sessions.

If we calculate a chi-squared statistic for the 2x2 table (Table 4) we get a chi-squared value of 34.84, which indicates a probability less than 0.001 of this outcome occurring at random.

The affective remarks are also very different in tone. There are no examples of apologies and contemplation (the most common categories in human sessions). 4 of the inputs to the computer are actively hostile, a category that does not appear in human sessions. 10 of the inputs are deliberate nonsense, which also does not appear in human sessions. In addition to the 20 examples listed in Table 3, there are two sequences in which the students are clearly testing the system; these we judged as rational experiment, rather than affect. One reason for these differences, we believe, is that we added a large number of "Why" questions, with the goal of obtaining longer, more

thoughtful responses from the students and some students did not choose to answer these questions. We also received a lot more null responses (no text, just a carriage returns) than we had seen before.

Sess	CIRCSIM-Tutor Question	Student Answer
M52	Why did you predict that IS would not change?	0
M52	Why did you predict that IS would increase?	1
M52	What does the baroreceptor reflex do?	1
M52	Can you explain why HR did not change?	no
M59	What does the baroreceptor reflex do?	nothing
M65	Can you explain why HR did not change?	+
M67	Why did you enter 'no change' for TPR?	BC
T48	Why did you enter 'no change' for TPR?	you know why
T48	Can you explain why HR did not change?	yes, i can.
T48	Why is MAP still decreased?	I don't want to tell you.
T56	Why did MAP change in the manner that you predicted?	[Student copies an earlier CST response]
T60	Why did MAP change in the manner that you predicted?	In other words, Nikie knows all
T60	Why did you predict that IS would not change?	it
T65	CO decreased in DR and increased in RR. Why did you predict that it would decrease in SS?	dr
T74	Why is MAP still decreased?	blalaal
T76	Why did you enter 'no change' for TPR?	the TPR can
T79	Why did you predict that IS would not change?	hatever
T81	Why is MAP still decreased?	asdf
T81	What does the baroreceptor reflex do?	t
T81	Why did you enter 'no change' for TPR?	Nimish said so.

Table 3: Possibly Affective Student Responses to CIRCSIM-TUTOR in November 2002

Sessions with/without Emotion	Human Tutors	Machine 2002	Totals
Emotion	22	12	34
No Emotion	3	54	57
Totals	25	66	91

Table 4: Sessions with and without Emotion (chi-squared = 34.8, $p<0.001$).

6. Implications for CIRCSIM-Tutor

Our transcripts do not conform to the Media Equation (Reeves & Nass, 1996). Students do not respond to CIRCSIM-Tutor in the same way that they respond to human tutors in keyboard-to-keyboard mode. They express much less emotion to the system than to human tutors and what emotion they do express is much less friendly. A comparison of hedges in these two sets of sessions showed that students hedge constantly to human tutors, but almost never to CIRCSIM-Tutor (Bhatt, Evens, & Argamon, 2004). Some of these differences may be due to the fact that CIRCSIM-Tutor does not begin by greeting the student and otherwise observe the social norms. Our expert tutors and colleagues at Rush Medical College felt that this approach was bogus, but now that it has become the norm for interactive software as well, we wonder if we should try to persuade them to reconsider. The work of Pon-Barry et al. (2006) suggests that it may be easier to detect emotion in spoken tutoring sessions.

One of the referees suggested that we look up the literature on threat analysis and we have added this project to our list of future research, although no student has yet threatened harm to us or our software. We are also considering redoing our markup using the approach in (Wiebe, Wilson, & Cardie, 2005).

7. Conclusion

We have transcripts of 75 human tutoring sessions, each approximately one hour long, available with and without our markup. We also have transcripts of machine tutor sessions with three different versions of CIRCSIM-Tutor (38 from 1998 plus 35 from 1999 plus 66 from 2002, making 139, all told). Although the machine tutor is modeled on the expert human tutors it does not have the same wit, charm, and natural language processing capability. Our machine tutor sessions contain fewer examples of overt emotion; the machine tutor sessions also involve more hostility and deliberate nonsense input. (Sometimes we find it hard to distinguish between emotion and attempts to test the system).

Please email the second author if you wish copies of any or all of these transcripts. Professor Joel Michael is happy to allow anyone to use these transcripts so long as acknowledgement is made to him and to Rush Medical College.

8. Acknowledgements

This work would have been impossible without the expert tutors, Joel Michael and the late Allen Rovick, both Professors of Physiology at Rush Medical College, and their students, and without the advice of Dr. Susan Chipman, expert in tutoring and tutoring systems. This work was partially supported by the Cognitive Science Program, Office of Naval Research under Grants No. N00014-94-1-0338 and N00014-02-1-0442, to Illinois Institute of Technology. The content does not reflect the position or policy of the government and no official endorsement should be inferred.

9. References

Bhatt, K., Evens, M.W., & Argamon, S. (2004). Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In K. Forbus, D. Gentner, & T. Regier (Eds.) *Proceedings of the Cognitive Science Society, COGSCI 2004*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc, pp. 114-119.

Evens, M.W., & Michael, J.A. (2006). *One-on-One Tutoring by Humans and Machines.* Lawrence Erlbaum Associates, Mahwah, NJ.

Fox, B. (1993). *The Human Tutorial Dialogue Project.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Glass, M.S. (1999). Broadening Input Understanding in a Language-Based Intelligent Tutoring System. Ph.D. Dissertation, Department of Computer Science, Illinois Institute of Technology.

Glass, M.S., & Evens, M.W. (2008). Extracting information from natural language input to an intelligent tutoring system. *Far Eastern Journal of Experimental and Theoretical Artificial Intelligence*, 1(2), to appear.

Kim, J.H., Freedman, R., Glass, M.S., & Evens, M.W. (2006). Annotation of tutorial goals for natural language generation. *Discourse Processes*, 42(1), 37-74.

Li, J., Seu, J.H., Evens, M.W., Michael, J.A., & Rovick, A.A. (1992). Computer Dialogue System (CDS): A system for capturing computer-mediated dialogues. *Behavior Research Methods, Instruments, and Computers (Journal of the Psychonomic Society)*, 24(4), pp. 535-540.

Michael, J.A., Rovick, A.A., Glass, M.S., Zhou, Y., and Evens, M.W. (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments*, 11(3), pp. 233-262.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16(2), pp. 171-194.

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge, UK: Cambridge University Press.

Shah, F., Evens, M.W., Michael, J.A., & Rovick, A.A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions, *Discourse Processes*, *33*(1), pp. 23-52.

Thompson, B.H. (1980). Linguistic analysis of natural language communication with computers. *Proceedings of the 8th International Conference on Computational Linguistics(COLING 80)*, Tokyo, Japan, np.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2-3), pp.165-210.

Vocal expression in spontaneous and experimentally induced affective speech: Acoustic correlates of anxiety, irritation and resignation

Petri Laukka[1], Kjell Elenius[2], Mats Fredrikson[1], Tomas Furmark[1], & Daniel Neiberg[2]
[1] Department of Psychology, Uppsala University, Uppsala, Sweden
[2] Centre for Speech Technology, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
E-mail: petri.laukka@psyk.uu.se

Abstract

We present two studies on authentic vocal affect expressions. In Study 1, the speech of social phobics was recorded in an anxiogenic public speaking task both before and after treatment. In Study 2, the speech material was collected from real life human-computer interactions. All speech samples were acoustically analyzed and subjected to listening tests. Results from Study 1 showed that a decrease in experienced state anxiety after treatment was accompanied by corresponding decreases in a) several acoustic parameters (i.e., mean and maximum F0, proportion of high-frequency components in the energy spectrum, and proportion of silent pauses), and b) listeners' perceived level of nervousness. Both speakers' self-ratings of state anxiety and listeners' ratings of perceived nervousness were further correlated with similar acoustic parameters. Results from Study 2 revealed that mean and maximum F0, mean voice intensity and H1-H2 was higher for speech perceived as irritated than for speech perceived as neutral. Also, speech perceived as resigned had lower mean and maximum F0, and mean voice intensity than neutral speech. Listeners' ratings of irritation, resignation and emotion intensity were further correlated with several acoustic parameters. The results complement earlier studies on vocal affect expression which have been conducted on posed, rather than authentic, emotional speech.

1. Background

Recent reviews of studies of vocal expression have shown that discrete emotions like anger, fear, joy, and sadness can be accurately communicated, also cross-culturally, and that each emotion is associated with relatively distinct acoustic characteristics (e.g., Juslin & Laukka, 2003; Laukka, 2008). However, the majority of previous research is subject to two major limitations – a) they have mainly been conducted on posed expressions and b) have limited their choice of emotion categories to full-blown basic emotions. It can reasonably be argued that posed expressions must be relatively similar to naturally occurring expressions in order for communication to be successful (Davitz, 1964), but posed expressions may also be exaggerated and more intense than authentic, everyday expressions (Scherer, 1986). Also, studies on spontaneous affect expression in everyday speech have reported that clear expressions of basic emotions are rarely found in normal day-to-day conversations, whereas expressions of milder and more subtle affective states occur more frequently (e.g., Campbell, 2005; Cowie & Cornelius, 2003; Devillers et al., 2005).

Researchers have attempted to study authentic vocal expressions in a number of ways. For instance, various affect induction methods have been applied in order to study the effects of the manipulation on the voice (e.g., Aubergé et al., 2006; Bachorowski & Owren, 1995; Bonner, 1943; Johnstone et al., 2005). Another line of research has investigated spontaneous emotional speech from real-life conversations (e.g., Devillers et al., 2005; Eldred & Price, 1958; Greasley et al., 2000; Lee & Narayanan, 2005; Litman & Forbes-Riley, 2006). These kinds of investigations are valuable, but also have limitations. For one thing, it is difficult to induce strong and well-differentiated emotional reactions in laboratory settings, which makes the study of intense emotional states difficult. Further, the study of real-life conversations is made complicated by the fact that one rarely has any control over what emotions, if any, the speakers actually were experiencing.

2. Aims

We present two studies on authentic vocal expressions which complement earlier research. More specifically, we explored the expression of non-basic emotions (anxiety, irritation, and resignation) using two different methods of collecting affective speech. In Study 1, the speech of social phobics was recorded during an anxiogenic public speaking task both before and after treatment. In Study 2, the speech material was collected from real life human-computer interactions.

3. Method

3.1 Study 1

The speech of patients with social phobia (N = 71, 45 women, 26 men, mean age = 35 years) was recorded in an anxiety provoking situation (i.e., giving a public speech) before and after pharmacological treatment, using data from a large ongoing project on treatment for social phobia conducted at Uppsala University. The patients were divided into responders and non-responders based on the effects of the treatment. Responders showed significantly less anxiety at post-treatment than at pre-treatment when compared to non-responders who served as a control group. Changes in self-reported state anxiety from pre- to post-treatment were evaluated using the Spielberger State-Trait Anxiety Inventory (Spielberger et al., 1970). The speech samples were

further analyzed regarding a number of acoustic parameters, and subjected to listening tests.

The recordings were made while the patients took part in a PET assessment and lay in the scanner. The patients were instructed to give a 2-minute speech about a vacation or travel experience, which was performed in the presence of a silently observing audience and the patients were instructed to observe the audience in order to increase observational anxiety. The first 10 seconds of speech from each subject at both recording occasions was subjected to acoustic analyses conducted with *Praat* (Boersma & Weenink, 2007).

The speech samples were further subjected to listening tests to test whether listeners could accurately perceive the speakers' level of anxiety. Sixteen listeners (8 men, 8 women, mean age = 24 years) judged the nervousness of all speech stimuli on a scale from 0 (not nervous at all) to 10 (very nervous). Before being entered into the listening test, the speech stimuli were content-masked by low-pass filtering. This procedure eliminates phonetic information and renders the speech unintelligible and sounding muffled. Nevertheless, affective information transmitted by F0, voice intensity and temporal aspects of speech is largely preserved. For a fuller description of the speech corpus used in Study 1, the reader is referred to Laukka et al. (in press).

3.2 Study 2

The speech material used was recorded from real life voice controlled telephone services by the Swedish company Voice Provider. The original database consisted of 61,078 utterances; mainly brief commands such as "yes" and "no", but also short sentences. All utterances were classified as neutral, emphasized or negative by a senior voice researcher. The majority of utterances were neutral, but the negative utterances included both irritated and resigned speech. Parts of this corpus have been used in prior studies on the automatic detection of affect from speech (e.g., Neiberg et al., 2006).

For our purposes we made a further selection of utterances from the Voice Provider database with the constraint that we should have at least one neutral and one affective utterance from each included speaker, in order to allow for within-subjects analyses. A further constraint was that the utterances should be of sufficient recording quality for acoustic analysis, and not contain any truncations, repetitions or other problems. The selected 200 utterances came from 64 different speakers and each speaker contributed between 2 to 6 utterances. Again, the speech samples were acoustically analyzed using Praat. Preliminary analyses of the acoustical characteristics of this selection of utterances have been reported in Forsell et al. (2007).

Twenty listeners (7 women, 13 men, mean age = 29.5 years) were asked to rate all speech samples on each of the following scales: irritation/anger, resignation/sadness, neutral, and level of emotion intensity. All scales ranged from 0 (not perceived at all) to 7 (very clearly perceived).

4. Results

Different sets of acoustic cues were analyzed in the two studies. Therefore we limit the presentation to cues which were common to both studies (e.g., pitch, voice intensity, spectral energy distribution, and temporal aspects of speech). For Study 1, we first wanted to examine the effect of anxiety on the acoustic measures. Because the decrease in self-reported anxiety from pre- to post-treatment was significantly larger for responders than for non-responders, we conducted planned comparisons (*t*-tests) between responders and non-responders for all acoustic cues. The analyses were conducted on the change scores from baseline to post-trial. To calculate the change scores for the voice cues, we subtracted the post-trial value for each cue from the baseline value for the same cue for each patient (i.e., we utilized a within-persons design). Thus a large change score indicated a large decrease from pre- to post-treatment. The results from the planned comparisons confirmed that the differences were significantly larger for responders than for non-responders for mean F0 (t_{69} = 2.52, $p < .05$), maximum F0 (t_{69} = 2.48, $p < .05$), and percentage of silence (t_{69} = 4.32, $p < .001$). In other words, a decrease in experienced anxiety was accompanied by corresponding decreases for the above acoustic cues. Additionally, we found a significant difference in change score for proportion of high-frequency energy (measured as the proportion of spectral energy above vs. below 1000 Hz [HF 1000]), indicating that proportion of high-frequency energy also decreased from pre- to post-treatment for responders, but not for non-responders (t_{45} = 3.45, $p < .01$).

We also wanted to relate the acoustic measurements with the listeners' perceptions of nervousness, because the correlations between listeners' ratings and voice cues give clues about what cues listeners use when making inferences about speakers' affective states. To that end we computed the correlations (Pearson *r*) between changes (pre- to post-treatment) in voice cues and listeners' mean ratings of nervousness, see Table 1. Changes in several voice cues were significantly correlated with changes in perceived nervousness. For example, decreases in maximum F0 (r = .24) and percentage of silence (r = .29) were associated with a decrease in perceived nervousness. Also, an increase in mean voice intensity was negatively correlated with an increase in perceived nervousness (r = -.35, all p's < .05).

For Study 2, we also first wanted to investigate acoustic differences between affective and neutral utterances. Therefore we first calculated the mean values for each acoustic measure (across all speech samples that were rated as neutral, irritated, or resigned by the 20 listeners) for each speaker. Then we investigated whether the mean

differences between neutral and irritated/resigned utterances were significant using within-groups *t*-tests. The tests revealed that mean F0 (t_{25} = 4.88, p < .0001) and the 5th quantile of F0 (F0 Q5; t_{25} = 3.22, p < .01) were significantly higher for irritated compared to neutral speech. The opposite results were obtained for resigned speech: both mean F0 (t_{16} = -2.24, p < .05), and F0 Q5 (t_{16} = -2.31, p < .05), were significantly lower than for neutral speech. Irritated speech further had higher mean voice intensity than neutral speech (t_{25} = 3.29, p < .01), whereas resigned speech had lower mean voice intensity than neutral speech (t_{16} = -2.57, p < .05). Also, irritated speech had higher H1-H2 (a spectral correlate of open quotient) than had neutral speech (t_{25} = 2.12, p < .05). Resigned speech additionally had longer mean syllable duration (i.e. slower speech rate) than neutral speech (t_{16} = 2.51, p < .05).

Voice cue	N	STAI-S	Nervousness
F0 M	71	.25 *	.16 ns
F0 SD	71	-.14 ns	-.06 ns
F0 max	71	.26 *	.24 *
Voice intensity M	47	-.21 ns	-.35 *
High frequency energy [HF 1000]	47	.35 *	.27 ns
Speech rate [syllables/s]	71	.18 ns	.02 ns
% Silence	71	.37 ***	.29 *

Note. * p < .05, *** p < .001

Table 1: Correlations (Pearson *r*) between changes (from baseline to post-trial) in voice cues and changes in a) speakers' self reports of anxiety (STAI-S) and b) listeners' ratings of nervousness from Study 1.

Finally we wanted to investigate the associations between acoustic cues and the listeners' perception of irritation and resignation. For this purpose we calculated the correlations between acoustic cues and the listeners' mean ratings of irritation and resignation, see Table 2. We found significant positive correlations between irritation ratings and mean F0 (r = .19), mean voice intensity (r = .44), and mean syllable duration (r = .20; p's < .01). Conversely, resignation ratings were negatively correlated with F0 standard deviation (r = -.40), F0 Q5 (r = -.25), and mean voice intensity (r = -.46). As also shown in Table 2, listeners' ratings of neutral and emotion intensity were significantly correlated with several voice cues.

5. Discussion

The presents results complement earlier studies on vocal affect expression which have been conducted on posed, rather than authentic, emotional speech. The studies have several noteworthy features, the implications of which will be briefly discussed below. First, we utilized within subjects designs, where each speaker acted as his or her own control. Thus, we could control for individual differences in baselines of the acoustic cues, whereas a lot of previous research has failed to include adequate control conditions with which to compare affective speech.

Second, we used two different kinds of affective speech corpora. Unlike most previous studies on induced affective speech, in Study 1 we managed to induce relatively strong affect. Nevertheless, the effects of the induced anxiety on the acoustics of speech were only small to moderate. One possible explanation of this could be that the speakers tried to mask their expressions in order to keep up a non-nervous appearance. The possibility of expression suppression clearly presents a challenge for the study of authentic vocal expressions and should be directly addressed in future investigations. In Study 2 we instead used recordings of spontaneous speech. Therefore we did not have any direct control over what emotions the speakers were experiencing. However, because of the context where the recordings were made (i.e., human-computer interactions where the communication was not working) we are fairly confident that the speakers actually did experience irritation and resignation, respectively.

Voice cue	Irritation	Resignation	Neutral	Emotion intensity
F0 M	.19 **	-.12 ns	-.15 *	.21 **
F0 SD	-.03 ns	-.40 ***	.22 **	.00 ns
F0 Q5	.14 ns	-.25 ***	-.03 ns	.16 *
Voice intensity M	.44 ***	-.46 ***	-.12 ns	.47 ***
H1-H2	.01 ns	-.11 ns	.00 ns	.05 ns
Mean syllable duration	.20 **	.08 ns	-.22 **	.15 *
Mean duration of silence	.04 ns	.00 ns	-.03 ns	.00 ns

Note. * p < .05, ** p < .01, *** p < .001

Table 2: Correlations (Pearson *r*) between acoustic cues and listeners' mean ratings of irritation, resignation, neutral, and emotion intensity from Study 2.

Third, the present results can be compared to previous results obtained with posed expressions (e.g., Juslin & Laukka, 2003). In general, the present results obtained with authentic expressions are similar to previous results obtained with posed expressions; though the effect sizes for non-portrayed affective speech are generally lower than for portrayed expressions (see also Williams & Stevens, 1972). Further, Study 2 is the first study that we know of where ratings of emotion intensity have been collected from spontaneous speech. The acoustical correlates of emotion intensity from Study 2 (see Table 2) are also in accord with previous studies conducted on posed expressions (e.g., Laukka et al., 2005).

Fourth, in the present studies we investigated the acoustical correlates of anxiety, irritation, and resignation, whereas most previous studies have instead investigated the basic emotions fear, anger, and sadness. As it turned out, the acoustical correlates of anxiety, irritation, and resignation were very similar to those of fear, anger, and sadness, respectively, which supports the hypothesis that affect in speech may be coded in terms of broad emotion categories/families (Laukka, 2005).

Fifth, the results from Study 1 lend support to the often hypothesized coupling between experienced emotion and expressive behavior, though they also speak against a one-to-one relationship between experienced and expressed emotion.

To conclude, designing experimental studies of authentic affect expression is fraught with difficulties, especially concerning how to collect the expressions. However, we believe that such studies are necessary to yield a better understanding of emotional vocal production, and about how and when emotion experience and various aspects of vocal expression co-occur.

6. Acknowledgments

This research was supported by the Swedish Research Council and the Sasakawa Young Leaders Fellowship Fund (SYLFF) through grants to the first author. We like to express our gratitude to Mimmi Forsell and Inger Karlsson for their invaluable help in the collection and analysis of the data from Study 2 and to Fredrik Åhs and Clas Linnman for their assistance in the data collection for Study 1.

7. References

Aubergé, V., Audibert, N., & Rilliard, A. (2006). De E-Wiz à C-Clone: Recueil, modélisation et synthèse d'expressions authentiques. *Revue d'Intelligence Artificielle, 20,* 499-527.

Bachorowski, J.-A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustical properties of speech are associated with emotional intensity and context. *Psychological Science, 6,* 219-224.

Boersma, P., & Weenink, D. (2007). Praat: Doing phonetics by computer (Version 4.6.12) [Computer program].

Bonner, M. R. (1943). Changes in the speech pattern under emotional tension. *American Journal of Psychology, 56,* 262-273.

Campbell, N. 2005. Getting to the heart of the matter: Speech as the expression of affect; rather than just text or language. *Language Resources and Evaluation, 39,* 109-118.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication, 40,* 5-32.

Davitz, J. R. (Ed.). (1964). *The communication of emotional meaning.* New York: McGraw-Hill.

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks, 18,* 407-422.

Eldred, S. H., & Price, D. B. (1958). A linguistic evaluation of feeling states in psychotherapy. *Psychiatry, 21,* 115-121.

Forsell, M., Elenius, K., & Laukka, P. (2007). Acoustic correlates of frustration in spontaneous speech. *Speech, Music and Hearing. Quarterly Progress and Status Report, 50,* 37-40.

Greasley, P., Sherrard, C., & Waterman, M. (2000). Emotion in language and speech: Methodological issues in naturalistic settings. *Language and Speech, 43,* 355-375.

Johnstone, T., van Reekum, C. M., Hird, K., Kirsner, K., & Scherer, K. R. (2005). Affective speech elicited with a computer game. *Emotion, 5,* 513-518.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129,* 770-814.

Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion, 5,* 277-295.

Laukka, P. (2008). Research on vocal expression of emotion: State of the art and future directions. In K. Izdebski (Ed.), *Emotions in the human voice. Vol 1. Foundations* (pp. 153-169). San Diego, CA: Plural Publishing.

Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion, 19,* 633-653.

Laukka, P., Linnman, C., Åhs, F., Pissiota, A., Frans, Ö., Faria, V., Michelgård, Å., Appel, L., Fredrikson, M., & Furmark, T. (in press). In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *Journal of Nonverbal Behavior.*

Lee, C. M., & Narayanan, S. (2005). Towards detecting emotion in spoken dialogs. *IEEE Transactions on Speech and Audio Processing, 13,* 293-303.

Litman, D. J., & Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication, 48,* 559-590.

Neiberg, D., Elenius, K., & Laskowski, K. 2006. Emotion recognition in spontaneous speech using GMMs. In *Proceedings of the 9th International Conference on Spoken Language Processing, Pittsburgh, PA.*

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143-165.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America, 52,* 1238-1250.

Multimodal records of driving influenced by induced emotion

Edelle McNahon[1], Roddy Cowie[1], Johannes Wagner [2], Elisabeth André [2]
Queen's University Belfast [1], University of Augsburg [2]
e-mail: e.m.mcmahon@qub.ac.uk, r.cowie@qub.ac.uk, johannes.wagner@student.uni-augsburg.de, Elisabeth.Andre@t-online.de

Abstract

Driving is a context that lends itself to the study of 'emotion in action'. A full-scale simulator has been used to study how prior induced emotion affects both standard indicator variables (vocal, facial and physiological) and the actions that a driver takes in a variety of contexts where emotion might be expected to be relevant. Induction uses a novel method which allows relatively strong emotions to be sustained and refreshed over a relatively long period. The contexts include a sequence of hazards; an opportunity to drive too fast; mental load; and sustained frustration. Analysis at an early stage, but emotion-related patterns, in both action and physiology, appear to emerge in poorly controlled responses to these contexts. It is intuitive that emotion may take that form, and the data illustrate the kind of resource needed to test whether it does.

1 Introduction

There is an increasingly widespread view that emotions are at root ways of relating to situations. They involve selective ways of perceiving situations (de Sousa 2004; Döring 2004; Deonna 2005) and distinctive ways of acting in and towards them (Frijda 1986, 2006); and they compete with other ways of perceiving and controlling action (Teasdale 1999, Cowie and Cornelius, 2003). The situation at the focus of an emotion may be part of the surrounding landscape, or part of the person's mindscape (for instance, an event long past, or anxiously anticipated). On that conception, the signs of emotion can be expected to pervade the person's 'action and interaction', in the phrase that was used in the HUMAINE network (http://emotion-research.net/projects/humaine/deliverables/d5a).

Until recently, most data collection reflected a very different conception of emotion. Broadly speaking, the records portrayed distinctive, short-lived episodes where a rush of feeling was accompanied by signs with a very specific link to that kind of feeling. The move away from that position has been gradual. There has been steady accumulation of data showing the various kinds of emotional colouring that are part of interaction (Cowie, Douglas-Cowie and Cox 2005). However, there is still relatively little data showing how emotion influences action and is expressed in the course of action. The omission is significant because it is bound up with our whole understanding of emotion. For instance, if we understand that people may read emotion from the way a person picks up a teacup, it should influence the way we think about signs of emotion in the voice. They may connect as indirectly to emotion as spilling the tea and ignoring the spill, not in the direct way that we tend to think a frown does, or a smile.

One of the problems is that action is so difficult to study over a period of any length. Driving is one of the scenarios where it seems possible to let people take action without feeling constrained, and still keep them in range of recording devices (cameras, microphones, and for that matter other body-worn sensors). As a result, it lends itself to a theoretically important kind of study.

Clearly, driving is also practically important, and a body of research on emotion and driving has accumulated for that reason. It has developed gradually. There is a long standing tradition of work that uses the term emotion, but which in practice focuses on stress and mental load (Fernandez and Picard 2003, McMahon, Cowie, Kasderidis, Taylor, and Kollias 2003). These are both easy to induce in driving tasks, and are both quite likely to precipitate emotion; nevertheless, they are not emotion in the ordinary sense. 'Road rage' attracts attention, but the research tends to treat it as something that lies outside the normal range emotion - a pathological symptom exhibited by maladjusted individuals (Galovski and Blanchard 2004). Some recent work has moved closer to investigation of emotion in the normal sense, but in a very a special context: studying how drivers react to in-car systems that express emotions like or unlike their own (Nass et al 2005). There is still very little work on the simple, central issue of the way emotion, in the ordinary sense, is reflected in the actions of a driver.

From a practical point of view, one of the central questions is whether easily obtained measures can identify the emotional states that are important for driving. Three main types of source have been considered. Voice has been studied extensively (Grimm et al 2007). Physiological methods are also commonly considered (Nasoz, Alvarez, Lisetti and Finkelstein 2004). Visual information, particularly information about about eye movements, has also been considered.

This paper reports studies that have generated a substantial body of data on the way emotion in a straightforward sense affects driving. The basic paradigm is as follows. Emotion is induced using a specially developed technique, EM3, to establish one of three states: angry, elated, or neutral. Participants then undertake driving tasks in a full scale simulator. The data recorded include physiological measures as

standard, and in some settings visual and auditory records, and records of performance on a secondary (attention) task; but also an instant-by-instant record of driving behaviour.

2 The techniques

2.1 Induction

EM3 was developed because existing techniques are limited in various ways. The methods most often used in driving research induce stress (or load), which is a different problem. Music changes driving behaviour, but not necessarily through its effects on emotion: its most obvious effect is to set a tempo which is difficult to break even if it is inappropriate. The emotion induced by standard techniques (such as the Velten procedure) tend to dissipate when participants have a task like driving to focus on. EM3 uses a Velten-like procedure followed by discussion of topics that it has been established in advance are highly emotive for the particular individual. The effects are long lasting and easy to refresh.

The core measures are standard physiological measures, ECG, GSR, Respiration, and skin temperature; and moment by moment records of the driver's road position, acceleration, and use of the controls (steering, throttle and brakes). These are supplemented by records of nearest approach to other vehicles and pedestrians, and of certain types of accidents (offroad excursions, speeding, crashes, etc.). Participants also used a standard instrument to self-rate their emotional state at key stages in each study.

2.2 Drives

Three basic type of drive were used to study key contexts in which emotion might be expected to affect action: responses when faced with sudden hazards; responses when a cognitive load is imposed; and responses to sustained obstacles.

Drives 1 and 2 presented participants with sudden and immediate hazards requiring rapid reactions.

A full intersection with signposts. Trees gradually disappear completely on approach to intersection. The participant has right-of-way, but a motorcyclist approaching from the left-hand side fails to yield. The motorcyclist is programmed so that if the driver keeps constant speed it will collide with the driver's vehicle.

Pedestrian Scenario 1. The approaches a roadside shop with a number of pedestrians standing outside the building. A bus is parked partially on the road so the driver will have to pull out onto the oncoming lane in order to pass the bus. There is light on-coming traffic. Pedestrians standing in front of the bus are hidden from the driver's view until he/she begins to overtake the bus. As the driver overtakes the bus two pedestrians and a dog suddenly walk out onto the road, and the driver must break suddenly to avoid a collision.

Obstacle: Abandoned Car. The driver approaches as sharp bend where a car is stopped partially in the driver's lane There is light on-coming traffic. The driver must brake, and then pull out to overtake the stalled vehicle.

Wrong lane vehicle. As the driver approaches a bend, an oncoming vehicle is approaching in the opposite lane, and a van is overtaking this vehicle in the driver's lane. As there is an approaching vehicle in the other lane, the driver has to pull onto the hard shoulder in order to avoid a collision.

Pedestrian Scenario2. The driver approaches a roadside shop, where a car is parked partially on the road, forcing the driver to pull out onto the oncoming lane in order to pass the parked car. There is light on-coming traffic. As the driver overtakes the car a child) suddenly runs out onto the road from the entrance of the shop. Because of the timing of the scenario (the pedestrian begins crossing just as the driver passes), the driver is less likely to collide with the pedestrian in this situation than in the previous pedestrian situation (in drive 1).

The cognitive load condition involved a relatively undemanding drive, during some periods of which participants had to respond to symbols overlaid on the 'windscreen' (roughly simulating a in-car information system). There were two periods of divided attention, each involving 15 symbol changes, roughly 600ft apart.

The sustained frustration condition was divided into periods of sustained hazard interspersed with a 2 mile period of relatively undemanding driving. In the first hazard, drivers quickly found themselves behind a slow-moving vehicle (a hearse) on a very winding stretch of road, with heavy oncoming traffic (72 vehicles over the 2 mile period), make overtaking the hearse very difficult. A 75mile stretch of straight road followed immediately after, giving the driver an opportunity to overtake the hearse if he/she chose. Drivers then encountered thick fog (visibility approx 50ft). The fog becomes visible in the distance, but from its edge to its thickest depth of opacity is a distance of only 150ft, so it seems to descend quite suddenly. Heavy oncoming traffic was also present.

2.3 Participants

There were 41 participants in the Sudden Hazards study (2 drives), and 45 in each of the Sustained Hazards and Cognitive Load condition (3 drives each). Drives lasted between 3 and 6 minutes (depending on drivers), making approximately 24 hours of data.

2.3 Processing

Data has been preprocessed to give accessible files. Driving data is structured into separate files for basic road position, episodes of hazardous behaviour, and outright incidents (crashes, etc). The raw physiological data has been analysed using the Augsburg BioSignal Toolbox, giving reliable measures of heart rate, breathing rate, etc. These measures can be related to each other and to records of the conditions to explore the way both overt action and visceral response relate to emotional state and external demands. Exactly as with other kinds of dataset, the records can be used as a basis for automatic recognition techniques aimed at identifying emotion on the basis of mode of action rather than simply facial or vocal signs. Recordings of facial and vocal signs are also available, though they have not been analysed. In contrast to most data sets, this one can also be used to look for predictors with an immediate practical applications, that is, predicting when a driver may be a risk to him/herself and others.

3 Illustrative data

We believe that this kind of data offers a rich picture of the way emotion affects and is reflected in multiple channels, in ways that depend profoundly on the context. Capturing the connections is a challenge for both psychology and machine learning, but both depend on the availability of appropriate data.

The figures below illustrate the kinds of connection that the data suggest may be present.

Figure 1 shows traces of key physiological indicators during the sudden hazards condition. The subject is angry. Various types of action, such as acceleration and braking, also depended on emotional state in that condition. What figure 1 shows is that in that condition, the physiological indices shifted substantially during the demanding events – suggesting that even if long-term anger does not change these indices in benign circumstances, it does dispose them to change markedly when sudden demands are imposed.

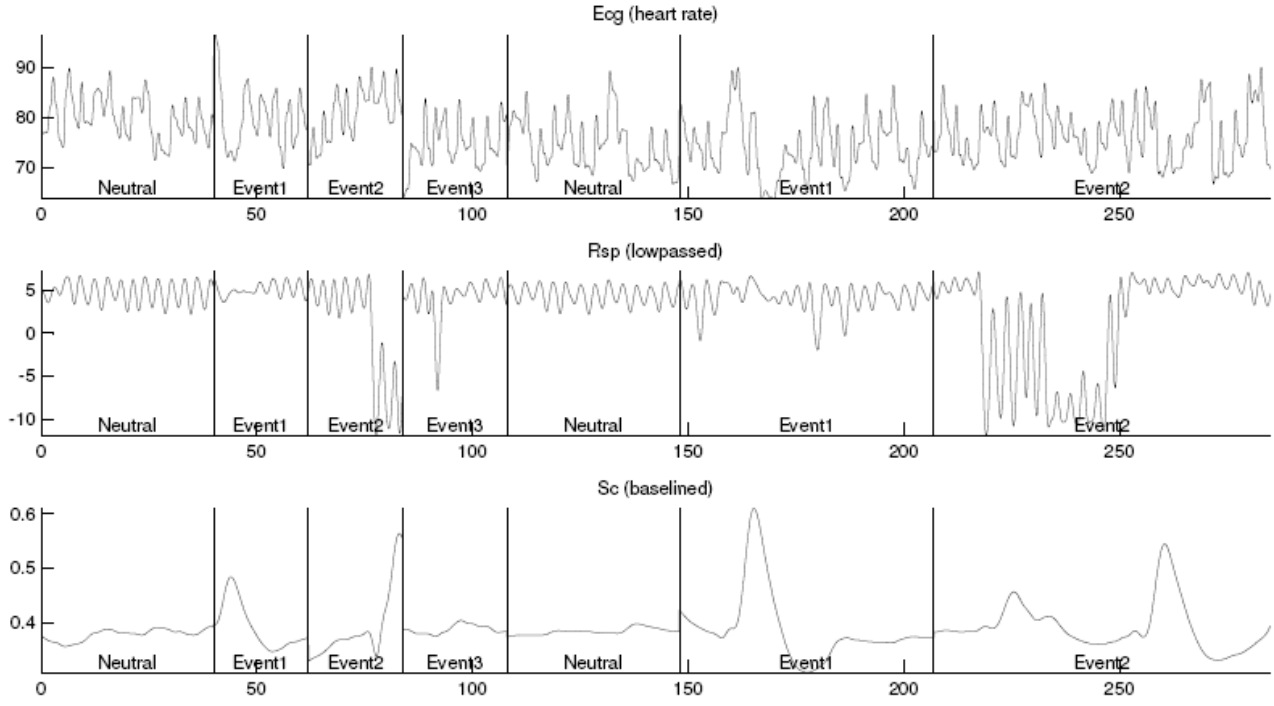


Figure 1: Processed physiological indicators in a sudden hazards drive

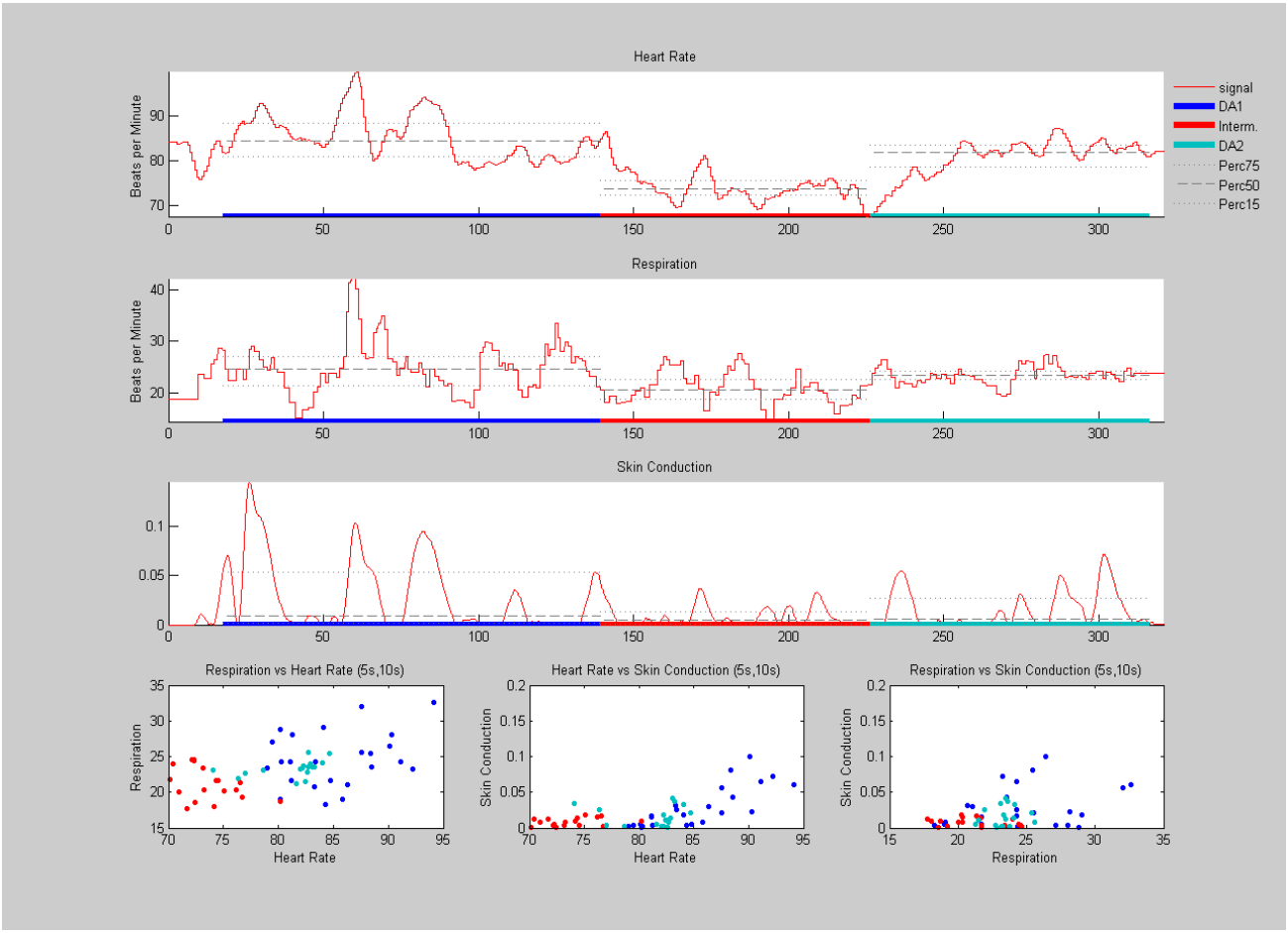


Figure 2 Processed biosignals in the two periods with cognitive load contrasted with a central undemanding period.

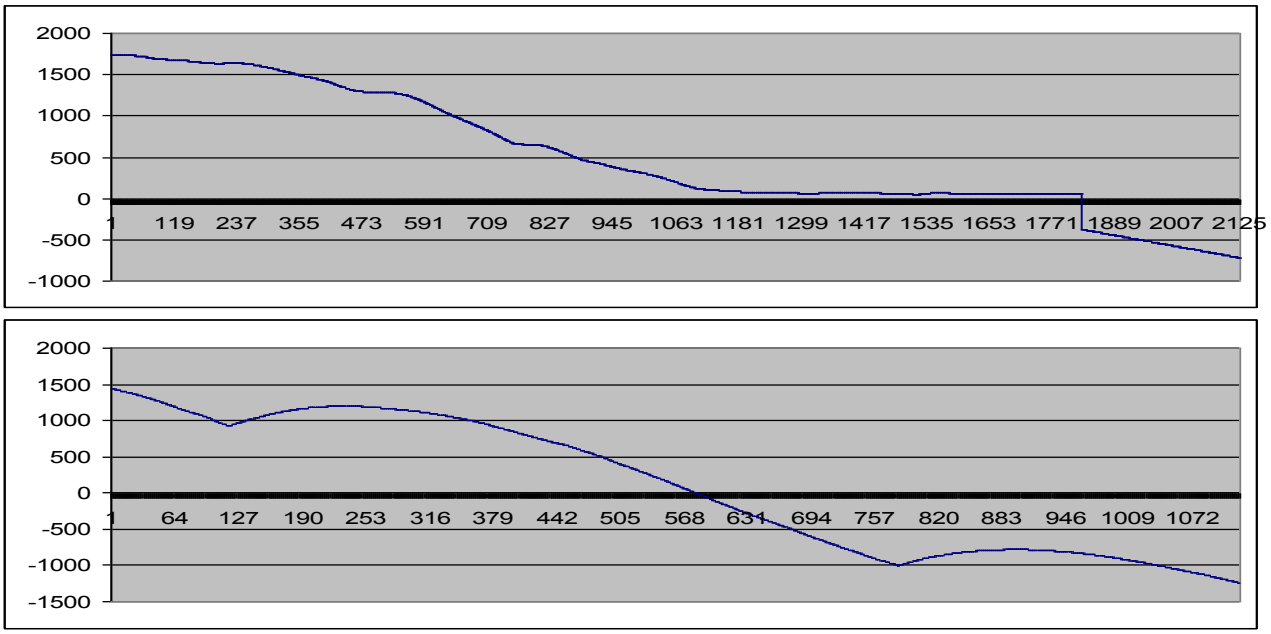


Figure 3 Distance to the slow car as a function of time in the least patient neutral driver (top panel) and an angry one (lower panel).

Figure 2 shows an angry participant during the cognitive load condition. Again, anger is reflected in over-reaction, shown in marked physiological changes between the load period and the undemanding drive.

Figure 3 shows the pattern of change in distance behind the slow-moving-vehicle for the least patient neutral participant and for an angry driver. It is not surprising that neutral drivers accept a period of delay behind the slow car, whereas angry drivers tend to overtake despite the risk. What is interesting is this database includes signs of emotion like that, which common-sense regards as highly symptomatic, but which present in few if any other databases.

4 Conclusion

It has been said that the job of philosophy is assembling reminders. The same is partly true of databases. The Belfast Driving Database assembles records that reflect ideas about emotion that are intuitive, but easily relegated to the margins in the process of designing procedures to elicit emotion and measure the ways it may appear.

Two central points are raised. One is that emotion is reflected not only in steady-state expressive behaviours, but also in behaviours that occur when certain circumstances arise. The second is that emotion may be reflected in reflected in a multiplicity of behaviours which have no specific relationship to emotion. What they reflect is the disruptive effect of emotion on systems that normally keep action within acceptable limits.

It is not yet certain what kinds of relationship the database will allow to be identified in a robust way. The description of results highlights relationships that are interesting, and consistent with intuition, but not proven. What matters is that the material documents action, physiological variables, and other signs, in a format that links the signs to records of the situations in which the signs occurred. That kind of evidence is needed to test reasonable but subtle intuitions about the way emotion is reflected in action and interaction.

5 References

Cowie R., & Cornelius, R.R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication Special Issue on Speech and Emotion: 40* (1–2), 5–32

Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). Beyond emotion archetypes: databases for emotion modelling using neural networks. *Neural Networks: 18*, 371-388.

Deonna, J.A. (2006). Emotion, perception, and perspective. *dialectica*, *60*, 29-46.

Döring, S.A. (2007). Seeing what to do: affective perception and rational motivation. *dialectica*, *61*, 363-394.

Fernandez R. and Picard R.W. (2003) *Modeling Drivers' Speech under Stress*. Speech Comm., 40:145-159

Frijda, N (1986) *The Emotions* Cambridge: Cambridge University Press

Frijda, N (2006) *The Laws of Emotion* Routledge

Galovski Tara E., & Blanchard Edward B. (2004) Road rage: a domain for psychological intervention? Aggression and Violent Behavior Volume 9, Issue 2, March-April, Pages 105-127

Grimm M, Kroschel K, Harris H, Nass C, Schuller B, Rigoll G & Moosmayr T (2007) On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving Proc Affective Computing and Intelligent Interaction 2007 Pages 126-138

McMahon, E. Cowie R., Kasderidis STaylor., J G., & Kollias S. *"What Chance that a DC Could Recognise Hazardous Mental States from Sensor Outputs?"*, Proceedings Disappearing Computer Tales, (DC Tales), Santorini, Jun 2003.

Nasoz F, Alvarez K, Lisetti CL, Finkelstein N (2004) Emotion recognition from physiological signals using wireless sensors for presence technologies Cognition, Technology & Work 6, 4-14

Nass C, Jonsson I-M, Harris H, Reaves, B Endo J, Brave S, Takayama L (2005) Improving automotive safety by pairing driver emotion and car voice emotion Proc. Conference on Human Factors in Computing Systems 2005 pp 1973 - 1976

de Sousa, R. (2004). Emotions – what I know, what I'd like to think I know, and what I'd like to think. In R.C. Solomon (Ed.), *Thinking about feeling* (pp. 61-75). Oxford & New York: Oxford University Press.

Teasdale, J.D., 1999. Multi-level theories of cognition–emotion relations. In:Dalgleish, T., Power, M. (Eds.), Handbook of Cognition and Emotion. John Wiley and Sons, Chichester, pp. 665–681.

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.

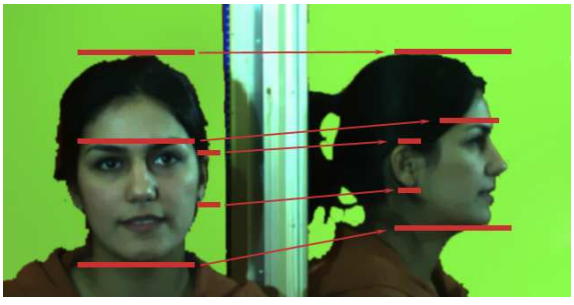


Figure 2: Sample frame with dual view.

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.

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).

Multi-modal emotion-related data collection within a virtual earthquake emulator

Dimitrios Ververidis, Irene Kotsia, Constantine Kotropoulos, and Ioannis Pitas
Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {jimver, ekotsia, costas, pitas}@aiia.csd.auth.gr

Abstract

The collection of emotion-related signals, such as face video sequences, speech utterances, galvanic skin response, and blood pressure from pupils in a virtual reality environment, when the pupils attempt to evacuate a school during an earthquake, is addressed in this paper. We assess whether pupil's spontaneous emotional state can be accurately recognized, using classifiers trained on elicited emotional speech and videos.

1. Introduction

A great expectation in human-centered computer interaction has been to exploit user's emotional state recognition as a feedback mechanism in order to adapt computer's response to user needs or preferences (Picard, 2000; Scherer, 2003; Ververidis and Kotropoulos, 2006a). In this paper, we report on an application-driven multi-modal emotion- related corpus collected in a virtual reality (VR) scenario, when pupils attempt to evacuate a school during an emulated earthquake. Several emotion-related bio-signals were recorded, while the pupils were immersed in the virtual earthquake environment. The data recorded were face videos, speech utterances, galvanic skin response for sweat indication, and blood pressure. Emotion recognition from facial videos (Kotsia and Pitas, 2007) as well as from speech (Ververidis and Kotropoulos, 2006b) have been thoroughly studied the past years. Sweat indicator and blood pressure signals have not been adequately studied yet, though related publications have been appeared and patents have been granted for their measurement. A wearable signal sampling unit with sensors mounted on the hand and the foot was developed in (Picard, 2000). Patents for sensors integrated with mouse, keyboard, and joystick have also been granted (Ark and Dryer, 2001).

The entire experiment was designed so that it provides objective evidence in order to evaluate the VR environment developed for training the pupils to cope with earthquakes, which frequently occur in Greece. In this paper, an assessment of the VR environment is presented, that is based on subjects' facial expressions, emotionally colored speech utterances, sweat indication, and the heart beat rate. An algorithm that recognizes the emotional state of a subject from speech is briefly discussed. To train the classifier, training data are used, whose elicited emotional state is known. Accordingly, the experiments are divided into two phases. In the first phase, the pupils learn how to express their emotions. In the second (or evaluation) phase, the pupils express their emotions during the emulated earthquake situation.

The outline of the paper is as follows. Data collection is described in Section 2. The classification of the collected facial videos is presented in Section 3. The classification of utterances into emotional state is accomplished via the Bayes classifier, which is described in Section 4. In Section 5, the galvanic skin response signal and the heart beat rate are analyzed. Finally, conclusions are drawn in Section 6.

2. Recording scenario

The VR environment emulates an earthquake occurring while the pupil is in a school classroom. Each pupil wears virtual reality glasses and sweat indicator and heart beat recording sensors prior to his immersion in the VR environment. Two microphones and a joystick, which is used for navigation within the VR environment, are placed nearby. A high resolution (near-field) camera is placed in front of the pupil in order to capture the head with high quality. Next to this camera a laptop exists that displays the VR environment the pupil sees. A second (distant) camera captures the entire experimental setup recording both the pupil and the virtual environment he/she is immersed in, in order to enable the annotation of the experimental recordings by psychologists and the synchronization of all input signals (video, audio, sweat indication, and heart beat signals). The experimental setup is depicted in Figure 1.

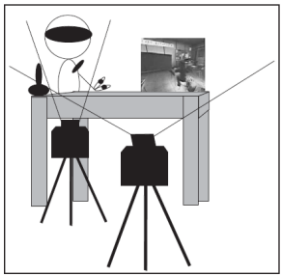


Figure 1: Experimental setup.

Regarding the experimental procedure, during the first phase, the pupils learn how to express their emotions. Episodes from several Greek movies were presented to the pupils. Each movie episode contains facial and speech expressions from an actor/actress colored by a certain emotion. By doing so, the pupils are familiarized with their role in the experiments. Two examples are depicted in Figure 2, where two pupil expressions are shown.

In detail, 14 pupils (5 boys and 9 girls) of age between 9 and 17 years were asked to express 13 utterances under 7 elicited emotional states. These utterances are used to


(a) (b)

Figure 2: Recording examples depicting (a) anger; (b) disgust.

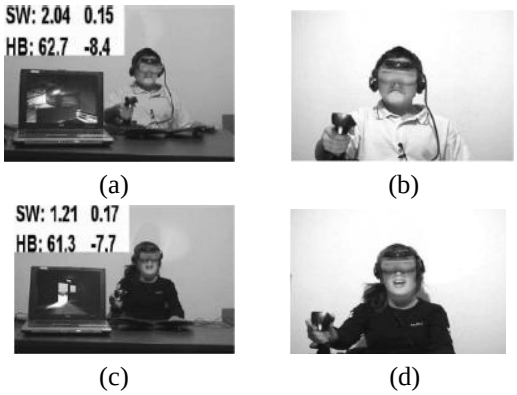
train the speech emotion recognition algorithm. The 7 emotional states of both the facial expressions and speech utterances are {anger, disgust, fear, happiness, neutral, sadness, and surprise}. The linguistic content of the utterances recorded during the first phase is described in Table 1.

Table 1: Linguistic content of utterances in Greek and their translation in English appearing inside parentheses.

1	*Αυλή* (Yard)
2	*Διάδρομος* (Corridor)
3	*Έξοδος* (Exit)
4	*Θρανίο* (Desk)
5	*Παιδιά* (Pupils)
6	*Παράθυρο* (Window)
7	*Πόρτα* (Door)
8	*Σχολείο* (School)
9	*Σεισμός* (Earthquake)
10	*Τάξη* (Classroom)
11	*Θα βγούμε στην αυλή αργά* (We shall go out to the yard slowly)
12	*Μπαίνω κάτω από το θρανίο* (I get underneath the desk)
13	*Περιμένω να σταματήσει ο σεισμός* (I wait until the earthquake stops)

The utterances collected in the first phase are 1396 (i.e. 14 (subjects) x 7 (states) x 13 (repetitions) plus some duplicates). In addition, a video capturing the facial expressions for each emotional state was recorded without any utterance by the pupil. During the first phase, 1 video sequence and two speech recordings (the first coming from the camera microphone and the second from a lavaliere microphone) were collected.

In the second phase, the pupils are immersed in a VR earthquake environment that consists of VR glasses and a joystick. The VR environment was developed on the top of the engine of the "Quake" game (Tarnanas et al., 2003). During the earthquake immersion, a virtual teacher (avatar) is giving instructions on how to cope with the situation, e.g. "Wait for the earthquake to stop" or "Proceed carefully to the exit". The objective is to assess pupil's emotional states within the VR environment. The following recordings were collected in the second phase: (i) 2 video sequences (one sequence capturing the facial expressions and another one recording the VR environment simultaneously with the pupils' expressions so that psychologists could evaluate the pupils' reactions); (ii) 3 speech recordings (2 recordings stem from the two cameras microphone and the third comes from a lavaliere



(a) (b)

(c) (d)

Figure 3: Captured snapshots from the distant camera for two subjects displaying sadness and fear (a) and (c) and the simultaneous high-resolution recordings of the face by the near camera (b) and (d).

microphone); (iii) 1 sweat indicator signal; (iv) 1 blood pressure signal.

The sweat indicator signal is the electrical conductivity between fingers (galvanic skin response, GSR) when a small electric current is applied. The blood pressure is measured by a plethysmograph, that is a pressure sensor positioned on a finger with a velcro strap. Through the blood pressure, one is able to measure heart beat rate by peak picking. Snapshots from the second phase are shown in Figure 3. In Figures 3(a) and 3(c), frames captured by the distant camera are shown. Sweat (SW) indication and heart beat (HB) rate at a certain time instant are overlaid in Figures 3(a) and 3(c). A laptop PC that shows exactly what the pupil sees was positioned near the pupil, so that the distant camera captures both the pupil and the VR scenes. Simultaneously, the second video camera records pupil's face, as shown in Figures 3(b) and 3(d). Technical details of the equipment are briefly summarized next. 2 PCs were used. The first PC was used to record the sweat signal, the blood pressure, and the speech from the lavaliere microphone. The second PC was running the VR environment. The following peripherals were used: 2 video cameras, 1 data recorder (IWorx-114), 1 pressure sensor (PT-100), 2 electrodes for measuring GSR (GSR-200), 1 sound sampling console (Behringer UB802), 1 condense microphone (AKG C417III), 1 joystick with force feedback, and 1 pair of VR glasses.

3. Classification of videos into emotional states during the immersion in VR

The method used to classify the facial expressions was the one proposed in (Kotsia and Pitas, 2007). However, due to the nature of the VR environment, only the facial expressions related to fear, sadness, and happiness were studied in the second (test) phase. Thus, the geometrical displacements of the deformed Candide grids were used as an input to a three-class Support Vector Machine (SVM) for facial expression recognition.

Fear is expected to appear most often during the experiments as it is the most common facial expression in

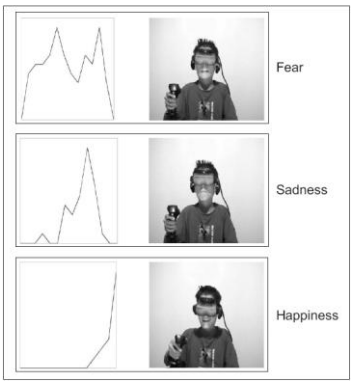


Figure 4:Evolution of anger, sadness, and surprise in time.

a case of an earthquake. Sadness is also expected as the pupil tends to be disappointed when obstacles prevent him/her to go out to the yard. When the pupil finally manages to get out (at the end of the immersion) he displays happiness.

These observations agree with the results collected when the system proposed in (Kotsia and Pitas, 2007) was used. An example of the evolution of the three aforementioned facial expressions in time is shown in Figure 4. As can be seen, fear is present during the entire video recording, apart from the beginning (when the pupil is in a neutral state) and the end (when the pupil is happy). In between, the intensity of fear decreases as sadness appears due to the several obstacles preventing the pupil's exit. Happiness appears only at the end of the recording, when the pupil manages to get out.

4. Classification of utterances into emotional states

The utterances collected during the first phase are used to train a speech emotion classifier. Speech from three emotional classes was used, namely, fear, happiness, and neutral. A set of 113 statistics of short-term pitch, energy, frequency contours is extracted, as in (Ververidis and Kotropoulos, 2006b). Each class-conditional probability density function of the extracted acoustic features is modeled by a multivariate Gaussian. Based on the aforementioned assumption, the Bayes classifier was designed. In order to find an unbiased estimate of the correct classification rate (CCR) admitted by the Bayes classifier, cross-validation is used, where 90% of the available utterances are exploited to train the classifier and the remaining 10% is used for classifier testing. Cross-validation is performed in a subject-independent manner. The average CCR for several cross-validation repetitions is used to estimate the CCR. The number of cross-validation repetitions for an accurate estimate of the average CCR is about 200 (Ververidis and Kotropoulos, 2006b). In order to avoid CCR deterioration, the Sequential Floating Forward Selection (SFFS) algorithm (Pudil et al., 1994) is used to select the feature subset that optimizes the CCR.

Two classification schemes are used, namely, the single-level scheme and the two-level one. In the single-level scheme, classification is performed in three classes. In the two-level scheme, two classifiers were employed. The first classifier is optimized by SFFS for separating {fear, happiness} vs. {neutral} states, and the second one is used for separating {fear} vs. {happiness}. The main idea behind the two-level scheme is that the acoustic features selected by SFFS in the first level are different than those selected by SFFS in the second level. The CCR achieved by the Bayes classifier with SFFS in the single-level scheme is 61.7%, when the random classification 33%. In Table 2, the classification rates among the three elicited states for each stimulus during training are shown. From the inspection of Table 2, it is deduced that the utterances expressed under the neutral state are easily recognized with a rate of 73.8%, whereas utterances colored by fear and happiness are recognized with a rates 58.7% and 52.5%, respectively. Anger and fear are often confused, due to their high arousal.

Table 2: Confusion matrix for the single-level scheme.

Stimuli/Response	Fear	Happiness	Neutral
Fear	58.7	19.9	21.4
Happiness	20.5	52.5	27.0
Neutral	14.3	11.9	73.8

The two-level classification scheme is depicted in Figure 5. The CCR achieved in the two-level scheme is 64.1%, i.e. there is an improvement of 2.4% against the single-level scheme. The confusion matrix of the two-level scheme is presented in Table 3. From the comparison between the confusion matrices in Tables 2 and 3, it is seen that the CCRs for fear and happiness are improved by 5%, whereas the CCR for the neutral state is reduced by 3%.

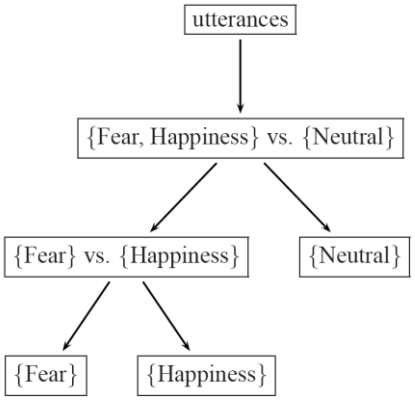


Figure 5: Classifying an utterance with the proposed two-level scheme.

Table 3: Confusion matrix for the two-level classification scheme.

Stimuli/Response	Fear	Happiness	Neutral
Fear	64.4	21.4	14.2
Happiness	25.7	57	17.3
Neutral	16.9	12.4	70.7

In the second phase, the Bayes classifier with the two-level classification scheme is used to classify another 155 utterances (disjoint to those used during the first

phase) into emotional states, which were expressed by the pupils during the VR immersion. The classification results are summarized in Table 4. From the 155 utterances, 91 utterances are classified into fear, 15 into happiness, and 49 into neutral state. Accordingly, it is deduced that pupils faced mostly fear during the VR immersion. This is an objective evidence demonstrating that the VR immersion level of pupils is large enough.

Table 4: Classification of utterances in the second phase.

Emotional state	Fear	Happiness	Neutral
Number of utterances	91	15	49
Percentage (%)	58.7	9.7	31.6

5. Sweat indication and heart beat rate

The sweat indication signal of 3 sample pupils among the 14 participated in the experiment is plotted in Figure 6. It is seen that the signal has many peaks and intense slopes in the first 50 sec, whereas a downward slope appears for the remaining 100 sec. This is due to the fact the virtual earthquake happens in the first 50 sec, and therefore kids become nervous. In the remaining 100 sec, kids are mostly focused on how to find the main exit of the virtual school, and accordingly they are more distracted and relaxed.

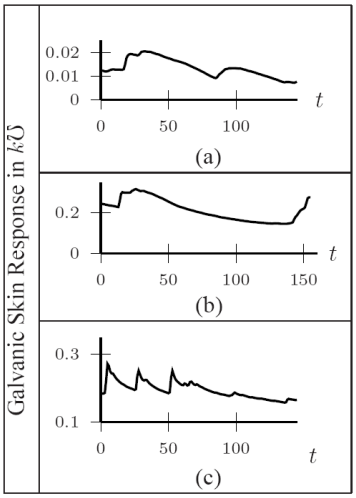


Figure 6: Sweat indication (GSR, electrical conductivity) plotted vs. time.

The heart beat rate signal of 3 sample pupils is plotted in Figure 7. From the inspection of these signals, a certain pattern can not be deduced. In Figure 7(a), an increasing slope of HB rate vs. time appears in the last 50 sec, when the pupil tries to find the school exit. In Figure 7(b), the pupil has approximately 100 pulses per minute without the HB rate function attaining any increasing or decreasing slopes. In Figure 7(c), the pupil's HB rate exhibits some peaks during the first 50 sec, and the HB rate function remains constant vs. time in the remaining 100 sec. The HB rate signal measured by the finger blood pressure is not so reliable as the sweat signal, because the pressure sensor is sensitive to the small movements of the finger.

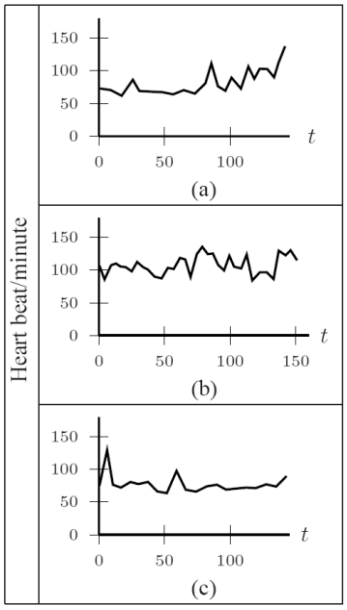


Figure 7: Heart beat rate as a function of time.

6. Conclusion

Emotion-related data have been recorded in the context of a VR earthquake scenario including facial video, emotional speech, and physiological signals. First results demonstrating the use of emotion recognition to assess the emotional state of pupils within the VR environment have been presented.

7. Acknowledgements

This work has been supported by project 01ED312 co-funded by the EU and the Greek Secretariat of Research and Technology (Hellenic Ministry of Development) of the Operational Program for Competitiveness within the 3rd Community Support Framework.

8. References

Ark, W., Dryer, C. (2001). Computer input device with biosensors for sensing user emotions. *US Patent 6190314*.

Kotsia, I., Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. Image Proc.*, 16(1), pp. 172-187.

Picard, R. (2000). *Affective Computing*. Cambridge: MIT Press.

Pudil, P., Novovicova, J., Kittler, J. (1994). Floating search methods in feature selection. *Pat. Rec. Let.*, 15, pp. 1119-1125.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Comm.*, 40, pp. 227-256.

Tarnanas, I., Tsoukalas, I., Stogiannidou, A. (2003). *Virtual Reality as a Psychosocial Coping Environment*. CA: Interactive Media Institute.

Ververidis, D., Kotropoulos, C. (2006a). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9), pp. 1162-1181.

Ververidis, D., Kotropoulos, C. (2006b). Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. In *Proc. European Signal Processing Conf. (EUSIPCO)*.

TRUE: an Online Testing Platform for Multimedia Evaluation

Santiago Planet, Ignasi Iriondo, Elisa Martínez, José A. Montero
GPMM – Grup de Recerca en Processament Multimodal
Enginyeria i Arquitectura La Salle, Ramon Llull University
Pg. Bonanova 8, 08022 Barcelona, Spain
E-mail: {splanet, iriondo, elisa, montero}@salle.url.edu

Abstract

TRUE (*Testing platfoRm for mUltimedia Evaluation*) is an online platform developed to create and perform subjective tests oriented to the evaluation of stimuli of different nature such as audio, video, graphics and text. Due to the high flexibility that the platform offers to researchers different kinds of tests can be carried out, such as emotion identification or quality assessment of synthesis systems, among others. The results can be used for different purposes depending on the research goals, e.g. to validate the emotional content of multimedia data of a corpus or to measure the expressivity of synthesized elements. TRUE involves all the stages related to the tests lifecycle, from their creation to the results retrieval, and allows the evaluators to answer the tests using any computer with an Internet connection. Making things easy for evaluators helps to minimize negative effects of fatigue, but also allows researchers to focus their efforts on the analysis of the tests results rather than on the supervision.

1. Introduction

Development of audiovisual corpora with authentic emotional content is one of the most challenging issues in the research on emotion and affect. Sources of emotional content can range from natural occurrences to acted performances (Campbell, 2000; Schröder, 2004), and a compromise between authenticity and recording quality should be considered. For this reason, tools to label and validate corpora content are required in order to guarantee a right performance in posterior use.

In general, two kinds of tests can be considered to face this goal: objective and subjective. The former does not require any kind of judgment from evaluators while the latter always involves the action of human raters. Subjective tests can be applied to: i) expressiveness validation of emotional audiovisual corpora, ii) labeling of corpora elements, including audiovisual, text and graphics resources, and iii) evaluation of synthesis systems, by rating individual stimuli or by comparing those synthesized by different techniques. Nevertheless, subjective tests can be useful in many other studies where a human criterion applied to the evaluation of data is required.

However, these subjective tests are usually designed to fit the particular features of the specific research and, in most cases, their designs are not reusable. Furthermore, evaluations tend to be time-consuming and tedious for users, whose fatigue could influence the results. In addition, achieving a high number of evaluators of heterogeneous profile tends to be difficult.

This paper describes TRUE, an online platform for designing and carrying out subjective tests that tries to solve the mentioned drawbacks. Section 2 describes TRUE features. Section 3 details the evolution of the platform from its creation to the current implementation. Section 4 explains the technology used for its development. Section 5 is devoted to published works that have used TRUE in order to gather subjective information. Finally, the conclusions and future work are presented in Section 6.

2. Description

TRUE copes with the requirements and the drawbacks mentioned in Section 1 by offering a tool that provides a single platform to design customized and reusable tests. Researchers can design the tests according to their needs and retrieve the results from the same tool, while evaluators can take the tests from any computer with an Internet connection. All the creation process of the tests and other management issues are carried out by means of web forms, as it is illustrated in Figure 1, where a form for the creation of a new test is shown.

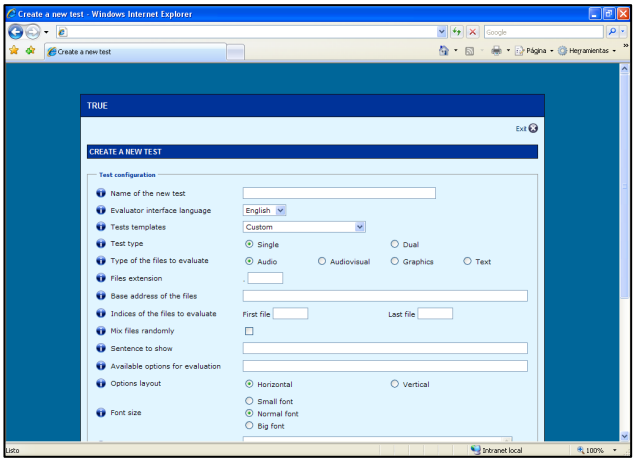


Figure 1: Form for the creation of a new test

The main goal of TRUE is to offer an interface for carrying out online tests. In this sense, TRUE gives a tool

for researchers to set up the tests allowing remote evaluators to rate the stimuli. Their answers are automatically stored in a database and can be recovered whenever the administrator requests them. Files to be tested can be stored in the same server than TRUE or in an external one because TRUE links these files from the test definition and shows the correct stimuli each time. This operation is shown in Figure 2.

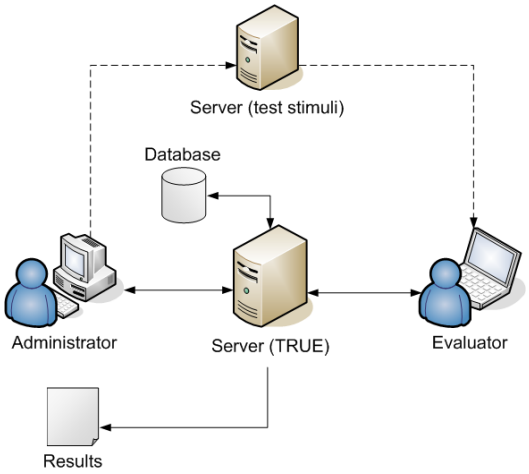


Figure 2: TRUE operation schema

It is important to highlight that different modalities (audio, video, graphics and text) can be used as stimuli and be suitably displayed in the user's web browser. Moreover, in order to avoid the negative effects of users' fatigue on the results, tests designed with the TRUE platform can be postponed and resumed. Another interesting feature is the inclusion of demonstrations at the beginning of a test which guides the answers from evaluators. A survey can be shown at the end of the test asking for user's profile and/or comments about the concluded test. Time spent on taking the test is also recorded.

With the aim of supplying standard tools to test designers, TRUE includes predefined templates for tests. These templates are related to MOS (Mean Opinion Score), CMOS (Comparison Mean Opinion Score) and DMOS (Degradation Mean Opinion Score) tests as defined by the International Telecommunication Union (1996). MOS tests are related to the assessment of perceived quality of various stimuli by means of a numerical indicator; CMOS tests perform a comparison between two stimuli; and DMOS tests are similar to CMOS but they measure the degradation in the stimulus quality when compared to another.

TRUE also allows the inclusion of plug-ins as templates. These plug-ins, such as Flash objects, can define specific tests – e.g. SAM (Self-Assessment Manikins) interface (Bradley & Lang, 1994), specifically oriented to emotion perception. The inclusion of plug-ins opens new ways of evaluating the stimuli far away from a forced answer test. In this sense, theories about representing emotions like points in small dimensional spaces instead of determining only a set of them, as Schröder (2004) stands, can be applied to online subjective tests. Figure 3 shows two different approaches for evaluating a stimulus: by means of a forced answer test by selecting an option from the radio buttons, or through the SAM based interface plug-in, where three dimensions of emotion are evaluated: activation, valence and dominance.

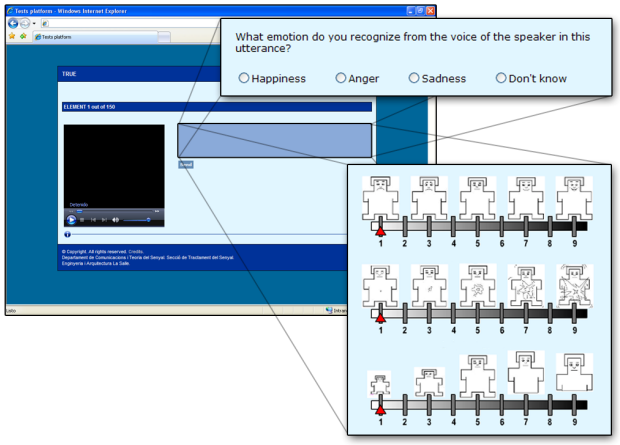


Figure 3: Evaluation by radio buttons and by means of a plug-in (SAM based interface, in this case)

TRUE differs from other test systems because it is not designed for a specific type of stimuli and it is very flexible. Unlike Irtel (n.d.) and NBS (n.d.), TRUE is not focused on a very wide range of areas of psychological research; it focuses on evaluation of corpora elements instead. Because of being focused on a specific area, and only oriented to online tests, its operation is easier than other systems like Empirisoft (2007), and permits the development of new evaluation tools for this research area. In addition to that, TRUE allows a broader audience than other tests that are usually conducted by means of paper forms or in a specific computer, like Schröder (2005). Possible inconsistencies caused by the use of an online system can be avoided by introducing some control elements. These elements can help to measure the rater's coherence.

3. Software evolution

The initial version of TRUE was designed to evaluate an emotional speech corpus in order to validate a study related to automatic emotion recognition (Planet, Morán & Formiga, 2006). The test consisted on a web with an embedded multimedia player where the user had to answer a question regarding the perceived emotion in the audio files. A set of radio buttons was used to show the studied emotional states (happiness, sadness, anger and neutral state). There were some sample audio files in the welcome page and a short survey at the end asking for age, sex and occupation of the evaluator along with his/her criteria when choosing the different options. The results were then compared with the ones from an automatic

classification following the approach of previous similar works as the described in Oudeyer (2003). A test alike to the one described above is shown in Figure 4.

Figure 4: A test for the evaluation of audio stimuli by means of radio buttons

According to the requirements of other research areas, and by tuning the configuration of the embedded player, the test was adapted to allow the evaluation of video files keeping the rest of the features unchanged. Another interesting feature for subjective tests is the comparison between different elements, for example in studies concerned to the evaluation of stimuli synthesized by different algorithms. This was put into service by permitting the inclusion of two embedded players and one or two questions related to the played files. The layout of the items could also be configured by the designer. Posterior versions included graphics and text as other possible evaluable elements. Figure 5 shows a dual test of graphical elements with horizontal layout.

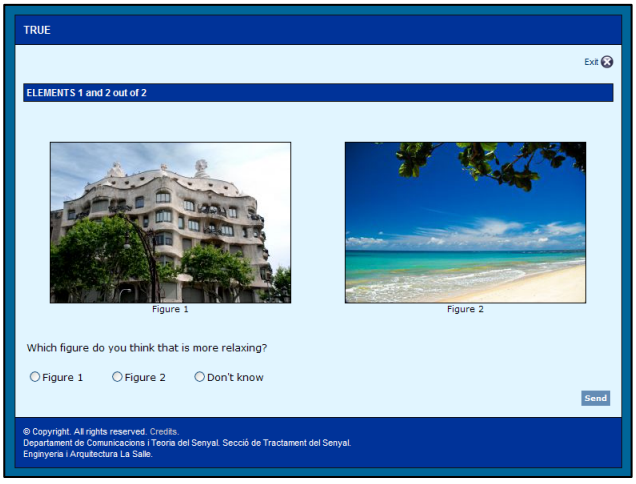


Figure 5: Test of two graphical elements with horizontal layout

Moreover, the welcome page was modified to offer different options: i) simple welcome without initial demonstration, ii) blind demonstration showing certain elements of the test with no further indications, and iii) full demonstration including some sample files with related comments. Each option can be fully customized by means of an embedded HTML editor in the creation web form or predefined templates can be chosen instead. The goal of these welcome pages is to give guidelines to the evaluators about the test, but also to familiarize them with the stimuli that they are going to rate in order to minimize errors during the evaluation. Figure 6 shows a welcome page with a full demonstration set up, which includes five audiovisual stimuli.

Figure 6: Full demonstration in the welcome page corresponding to an audiovisual test

The final survey can also be customized allowing the selection of the number and type of questions to be asked to the users. These questions can be text fields, text areas, radio buttons and lists. Figure 7 shows a final survey sample.

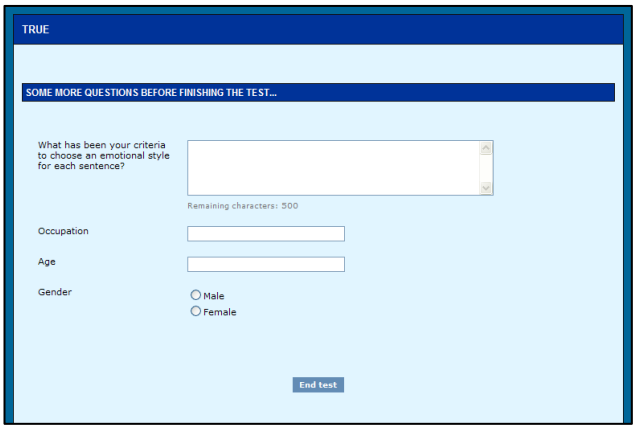


Figure 7: Example of survey at the end of a test

In all the cases, TRUE can be customized for different languages by installing the appropriate configuration files. These files are text documents with translations for the web elements and information related to the set up of the installed plug-ins.

4. Technology

TRUE is a web service implemented in Java©. It provides a tool for building and performing subjective tests, storing the evaluations of the different elements and offers management tools. Stimuli files can be placed in the same server where TRUE is installed or in another external one; this feature entails the storage capacity of the server does not need to be large. Flexibility of the platform is not only shown in the creation process but also in the reusability of previous tests and in the retrieval of results in different formats, like Microsoft Excel© or CSV (Comma Separated Values). These results, as other data related to the tests, are stored in a MySQL© database in the server and downloaded when the user requests this information; nevertheless, internal processes are transparent to end users.

5. Related studies

Several studies carried out within the authors' research group (GPMM) have used the TRUE platform. Concerning to emotion identification from speech, TRUE has been used to validate an emotional speech corpus for its use in an automatic emotional recognition experiment (Planet, Morán & Formiga, 2006). The goal of the study was to evaluate different data mining techniques comparing the performance of the algorithms and the recognition rate achieved by human evaluators.

In a broader sense, in three recent studies (Iriondo et al., 2007c; 2007b; 2007a) the objective was the automatic validation of a whole emotional speech corpus by mapping subjective criteria to automatic classification algorithms. The corpus consisted of 4638 utterances corresponding to five expressive styles: neutral-mature, joy-elation, sensual-sweet, aggressive-hard and sad-melancholic. Only 480 utterances were randomly chosen (96 per style) to be subjected to a forced answer test with the question: *What emotion do you recognize from the voice of the speaker in this utterance?* The possible answers were the five styles plus an extra option *Don't know / Another*, to avoid biasing the results because of confusing cases. The first of the studies revealed differences between classification errors made by the automatic algorithms and human evaluators although in both cases the percentages of identification were very high. The subsequent studies were focused on emulating these subjective results by mapping them to the automatic classification.

In Iriondo, Socoró & Alías (2007), tests are related to expressive speech synthesis, with the aim of measuring the quality of the synthesized utterances. In Gonzalvo et al. (2007a; 2007b), TRUE is used with the same goal. In the three cases, subjective viewpoint is measured by MOS tests.

Audiovisual analysis and synthesis has been covered in Sevillano, Melenchón & Socoró (2006) and Melenchón (2007).

TRUE can be helpful in a wide range of purposes. As an example, it has been successfully used in research about learning methodologies (Montero et al. 2007). In this work, the subjective tests made by TRUE collected expert knowledge from teachers, which was later modeled by a fuzzy logic system able to rate the teamwork performance of engineering students.

6. Conclusion and future work

This paper has presented TRUE, an online platform designed to develop subjective tests. Although TRUE's main goal was to help in studies related to validation of emotional speech corpora, its flexibility makes it useful for a wide range of stimuli sources and purposes, e.g. evaluation of audiovisual synthesis algorithms. Because of being an online platform, tests are very accessible for users. This is a great advantage for tests designers as this makes it possible to reach a broader audience keeping an easy way of developing their tests.

TRUE has proved to be a very helpful tool in different research fields. Authors wish it can ease other researchers work and invite them to download TRUE from the TRUE website [1]. Feedback on its use and suggestions for improvement will be appreciated. TRUE's design allows the inclusion of new features related to the kind of tests to be created. In these sense, different plug-ins are being developed to fit the requirements of different research areas. Moreover, any researcher can design specific plug-ins according to the guidelines provided in the development documentation and include them in their TRUE installation; it does not require any additional change in the platform.

As detailed in Section 5, TRUE has been used in many studies to gather a big amount of subjective data. In all these studies, a later stage consisted on a statistical analysis to obtain relevant conclusions using hypothesis tests such as ANOVA, t-student, Kolmogorov-Smirnov, etc. Current work is focused on embedding these tools in order to allow the users to ask for these statistical analyses during the result retrieval process.

7. Acknowledgements

This work has been partially supported by the Spanish Science and Education Ministry (CICYT TEC2006-08043/TCM).

8. References

Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25(1), 49--59.

Campbell, N. (2000). Databases of emotional speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 34--38.

Empirisoft (2007). Medialab and DirectRT Software for

[1] http://www.salle.url.edu/tsenyal/true

Psychology Experiments. Retrieved April 5, 2008, from http://www.empirisoft.com/medialab.aspx.

Gonzalvo, X., Iriondo, I., Socoró, J. C. & Monzo, C. (2007a). Mixing HMM-based spanish speech synthesis with a CBR for prosody estimation. In *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007. Lecture Notes in Computer Science*, 4885, pp. 78--75. Springer, Heidelberg.

Gonzalvo, X., Socoró, J. C., Iriondo, I. & Monzo, C. (2007b). Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castillian Spanish. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6)*, pp. 362--367, Bonn, Germany.

International Telecommunication Union. (1996). *Methods for subjective dedtermination of transmission quality. ITU-T Recommendation P.800.*

Iriondo, I., Planet, S., Alías, F., Socoró, J. C. & Martínez, E. (2007a). Validation of an expressive speech corpus by mapping automatic classification to subjective evaluation. In *Computational and Ambient Intelligence. 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings. Lecture Notes in Computer Science*, 4507, pp. 646--653. Springer, Heidelberg.

Iriondo, I., Planet, S., Alías, F., Socoró, J. C., Monzo, C. & Martínez, E. (2007b). Expressive speech corpus validation by mapping subjective perception to automatic classification based on prosody and voice quality. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007).* Saarbrücken, Germany.

Iriondo, I., Planet, S., Socoró, J. C. & Alías, F. (2007c). Objective and subjective evaluation of an expressive speech corpus. In *Advances in Nonlinear Speech Processing. International Conference on Non-Linear Speech Processing, NOLISP 2007, Paris, France, May 22-25, 2007. Lecture Notes in Computer Science*, 4885, pp. 86--94. Springer, Heidelberg.

Iriondo, I., Socoró, J. C. & Alías, F. (2007). Prosody modelling of Spanish for expressive speech synthesis. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 4, pp. 821--824. Honolulu, HI, USA.

Irtel, H. (n.d.). PXLab: The Psychological Experiments Laboratory. Retrieved April 5, 2008, from http://www.pxlab.de.

Melenchón, J. (2007). Síntesis Audiovisual Realista Personalizable. PhD Thesis . Enginyeria i Arquitectura La Salle, Universitat Ramon Llull.

Montero, J. A., Alías, F., Garriga, C., Vicent, L. & Iriondo, I. (2007). Assessing students' teamwork performance by means of fuzzy logic. In *Computational and Ambient Intelligence. 9th International Work-Conference on Artificial Neural Networks, IWANN 2007, San Sebastián, Spain, June 20-22, 2007. Proceedings*. *Lecture Notes in Computer Science*, 4507, pp. 383--390. Springer, Heidelberg.

NBS (n.d.). Auditory, Visual and Multi-modal Stimulus Delivery for Neuroscience. Retrieved April 5, 2008, from http://www.neurobs.com/presentation.

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2), 157--183, special issue on Affective Computing.

Planet, S., Morán, J. A. & Formiga, L. (2006). Reconocimiento de emociones basado en el análisis de la señal de voz parametrizada. In *Actas da 1a Conferência Ibérica de Sistemas e Tecnologias de Informação, Ofir, Portugal, 21 a 23 de Junho de 2006*, 2, pp. 837--854.

Schröder, M. (2004). Speech and Emotion Research: An Overview of Emotion Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD Thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University.

Schröder, M. (2005). RatingTest - Java software for designing and carrying out listening tests. Retrieved April 5, 2008, from http://ratingtest.sourceforge.net.

Sevillano, X., Melenchón, J. & Socoró, J. C. (2006). Análisis y síntesis audiovisual para interfaces multimodales ordenador-persona. In *Proceedings of the 7th Congreso Internacional de Interacción Persona-Ordenador (Interacción 2006)*. Puertollano, Ciudad Real.

Spanish Expressive Voices: corpus for emotion research in Spanish

R. Barra-Chicote[1], J.M. Montero[1], J. Macias-Guarasa[2], S.L. Lufti[1], J.M. Lucas[1], F. Fernandez-Martinez[1], L.F. Dharo[1], R. San-Segundo, J. Ferreiros[1], R. Cordoba[1], M. Pardo[1]

Universidad Politecnica de Madrid[1] Universidad Alcalá de Henares[2]

{barra, juancho, syaheerah, juanmak, efhes, lfdharo, lapiz, jfl, cordoba, pardo}@die.upm.es[1], macias@depeca.uah.es[2]

Abstract

A new emotional multimedia database has been recorded and aligned. The database comprises speech and video recordings of one actor and one actress simulating a neutral state and the Big Six emotions: happiness, sadness, anger, surprise, fear and disgust. Due to a careful design and its size (more than 100 minutes per emotion), the recorded database allows comprehensive studies on emotional speech synthesis, prosodic modelling, speech conversion, far-field speech recognition and speech and video-based emotion identification. The database has been automatically labelled for prosodic purposes (5% was manually revised). The whole database has been validated thorough objective and perceptual tests, achieving a validation score as high as 89%.

1. Introduction

Several multimedia corpora have recently been developed involving speech studies. However, some of them are limited to either the analysis of emotion expression, or detection.

Most of these corpora focus on meeting settings (Chen et al,2005) (Mana et al, 2007) and even then these studies are mostly focusing on multispeaker solutions, lip-reading correlating to human understanding, personality traits and social behaviours while discussing and interacting - and not particularly concentrating on the synthesis or detection on emotions per se. In addition, one of the main problems in speech recognition tasks is how to adapt classifiers to affective speech.

Several studies (Mana et al, 2007)(Castellano et al, 2007)(Sebe et al, 2005) has highlighted that an ideal system for automatic recognition of human affective information should be multimedia. This integration is also exhibited in psychology studies such as (Schrerer et al, 2007). However, the studies mentioned above introduce multimedia approaches that are limited to automatic affect sensing and not for the use of affective expressions such that those in text-to-speech systems.

The paper is organised as follows: previous considerations and description of target information is explained. Next, whole acquisition equipment and set-up is described. Finally, the evaluation of the corpus is presented and its labelling method explained.

2. Corpora Design

2.1. Previous Considerations

What makes SEV unique is that it is a combination of four Spanish previous corpora. Each of them designed to meet a particular goal and equipped with various underlying objectives of speech studies which are not just limited to detection of emotion, but also to acceptable synthesize of emotions (for expressions).

The acquisition of enough far-field emotional speech would give the possibility of the adaptation of beam-forming techniques to emotional speech that could perform the speech applications growing in human-machine interfaces where no close-talk speech signal is available.

The addition of video information could provide features that help to provide more specific information such as detection of level of intensity of a particular emotion or even make up for information lost due to corrupting influences in the audio content.

Several considerations like what kind of emotions would be recorded, what emotions should be recorded or who will be the speakers should be taken into account.

First discussion is the pros and cons of acted versus real emotions. As pointed in (Burkhardt et al., 2005), so-called full-blown emotions very rarely appear in real world and ethic problems appear when recording people experiences of real emotions. These things make almost impossible to work with real data and do it in a clean and high quality acquisition set-up.

The multimedia character of SEV makes more difficult to approach to real situations without losing quality research aspects. This fact makes a difficult task the expression of emotions simultaneously through speech and mimic (mainly facial features). This difficulty and one of the main purposes of the corpus, high quality emotional speech synthesis, are the main reasons why two professional actors (one male and one female) were selected.

It seems logical to use distinct terms when acted emotions are investigated. Due to this our SEV corpus focuses on discrete emotion instead of emotional states projected on "emotional dimensions" (PAD model (Schoder, 2004)). Looking for comparison between new

studies and previous works in our Group (Montero et al., 2002)(Barra-Chicote et al., 2006)(Barra-Chicote et al., 2007) we selected original emotions in SES corpus (Montero et al., 1998) (happiness, cold anger, surprise, sadness and neutral reference) and three new ones (fear, disgust and hot anger) in order to complete the group of basic emotions.

It has been pursued that SEV corpus would be well suited for use to analyse data in developing or testing automatic recognition systems or systems involving emotion synthesis and hopefully there is performance increasing in these two tasks.

2.2. Speech Content

The main purpose of 'near' talk speech recordings is emotional speech synthesis and recognition and tasks related to emotion identification. Three channels were recorded: a close talk headset-microphone, a lapel microphone and a desktop microphone.

Several length features of the corpora, as average length of words (< w >) or allophones (< A >), are presented in Table 1.

- **Emotional Level Corpus**

Fifteen reference sentences of SESII-A corpus were played by actors 4 times, incrementing gradually the emotional level (neutral, low, medium and high level).

- **Diphone concatenation synthesis corpus**

LOGATOMOS corpus is made of 570 logatomos within the main Spanish Di-phone distribution is covered. They were grouped into 114 utterances in order to provide the performance of the actors. Pauses between words were requested to them in the performance in order to be recorded as in an isolated way.

This corpus allows studying the impact and the viability of communicate affective content through voice by no semantic sense words. New voices for limited domain expressive synthesisers based on concatenative synthesis would be built.

- **Unit Selection synthesis corpus**

QUIJOTE is a corpus made of 100 utterances selected from the 1st part of the book Don Quijote de la Mancha and that respects the allophonic distribution of the book. This wide range of allophonic units allows synthesis by unit selection technique.

- **Prosody Modelling In SESII-B corpus**

Hot anger was additionally considered in order to evaluated different kinds of anger.

The 4 original paragraphs in SES (Montero et al., 1998) has been split into 84 sentences. PROSODIA corpus is made of 376 utterances divided into 5 sets. The main purpose of this corpus is to include rich prosody aspects that makes possible the study of prosody in speeches, interviews, short dialogues or question-answering situations.

2.3. Far Field Content

The main purpose of 'far' talk speech recordings is to evaluate the impact of affective speech capture in more realistic conditions (with microphones placed far away from the speakers), also in tasks related to speech recognition and emotion identification.

Two microphone arrays were used for recording: a linear harmonically spaced array composed of 12 microphones placed on the left wall, and a roughly squared microphone array composed of four microphones placed on two tables in front of the speaker.

Although the acoustic environment is controlled and reverberation is low, experiments on this data can lead to interesting results on emotional speech processing.

2.4. Video Content

Video information was also recorded for every utterance. The main purpose of this capture is allowing research on emotion detection using visual information, face tracking studies and the possibility of study specific head, body or arms behaviour that could be related to features such as intensity level in the recorded speech signals or give relevant information of each emotion played. Also, audio-visual sensor fusion for emotion identification and even affective speech recognition are devised as potential applications of this corpus.

Figure 1 shows some zoom examples of various emotions and Figure 4 presents the camcorder situation in the chamber.

Figure 1: Example frames of emotions

3. Recording Equipment and Setup

This section describes the equipment used to record the SEV database. The recording was controlled by an operator in a room next to the recording acoustic treated chamber. The operator controlled the recording

application and was able to talk to the speaker at any time. Recording equipment was able to synchronously record 20 audio channels and a video signal. Figure 2 shows the general architecture of the recording set-up.

3.1. Audio Recording Hardware

Every channel's audio is sampled at 48 kHz and samples are 24 bits long. Twenty channels consist of:

- Channel 1: A close-talk head-mounted Shure Microphone.
- Channel 2: An electroglotograph signal is recorded in order to get high-quality pitch marks. The actors were a necklace with two electrodes as shown in Figure 4.
- Channel 3: A desktop microphone.
- Channel 4: A lapel Shure Microphone.
- Channels 5-8: A quasi-squared array composed of 4 PZM microphones.
- Channels 9-20: A harmonically-spaced linear array composed of 11 microphones, plus an additional microphone (for z-axis position discrimination) located 30 cm above the linear array axis.

All audio sources are connected to 3 RME Octamic-D units, in charge of microphone pre-amplification and A/D conversion.

The channels devoted to the Sennheiser mics are phantom powered (using the provided phantom power from the microphone pre-amplifiers). The built-in high pass filter (LO CUT option) of the microphone pre-amplifiers is activated (80Hz, 18 dB/octave).

The 3 ADAT digital streams from the Octamics are optically linked to a RME HDSP 9652 acquisition PCI card installed into a dedicated Dual-Xeon PC running Windows XP. The recording room is acoustically treated to acoustically isolate the chamber from the outside and to reduce reverberation (although there are reflective surfaces inside).

3.2. Room Geometry and Microphones Localisation

In Figure 3 we show a wire-frame simulated version of the room, in which several room devices can be seen: the rectangular table in the middle of the room, the door, the window, the frame holding the linear microphone array, and the positions of the desktop microphones and the four ones composing the quasi-squared microphone array.

The MC-REC recording application shows the text and the specific emotion that should be played by the actor. Once the operator clicks the record button, the application indicates the actor to start by showing a red circle.

The video acquisition was carried out by using a dedicated GNU/Linux workstation, running as a video server. The camera was controlled using the dvgrab free software application, modified in order to be able to work in client-server mode. Files were recorded using *720x576* resolution and 25 frames per second.

The MC-REC application is also in charge of requesting video recording to the video server, with a simple protocol (TCP/IP socket based) specifying start/stop commands and filename options.

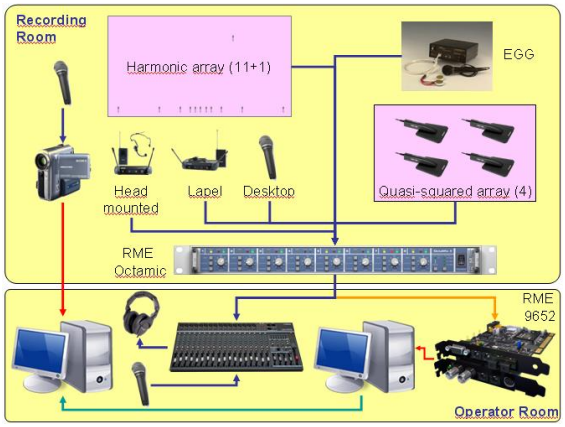


Figure 2: Schematic of recording setup

3.3. Cons aspects

The wide range speaking style for the acted emotions, made a difficult task to adjust mic recording levels in order to keep quality constant. Another problem due to the huge amount of recording sessions, was to keep the emotional patterns and emotional intensity constant during all sessions. The acquisition of SEV has taken more than 40 hours of recording sessions, distributed during one month (it was harmful for actors to play more than three hours a day).

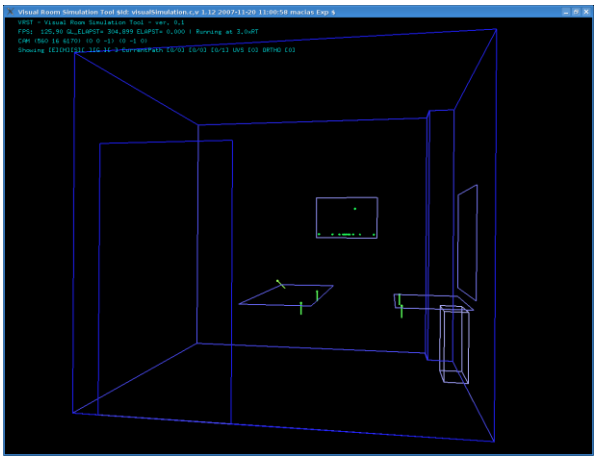


Figure 3: Wireframe simulation of recording room (3D view)

4. Evaluation

Close talk speech of SESII-B, QUIJOTE and PROSODIA corpus (3890 utterances) has been evaluated using a web interface. Six evaluators for each voice participated in the evaluation. They could hear each utterance as many times they need. Evaluators were asked for the emotion played on each utterance and its emotional level (choosing

between very low, low, normal, high or very high). 60% of utterances were labelled at least as a high level utterance.

Each utterance was evaluated at least by two people. The Pearson coefficient of identification rates between the six evaluators was 98%. A kappa factor of 100% was used in the validation. 89.6% of actress utterances and 84.3% of actor utterances were validated. Figure 5 plots emotion validation results.

Figure 4: Example of recording session

Whole database has been evaluated by an objective emotion identification experiment that leads 95% identification rate for both speakers. Automatic emotion identification was based on PLP speech features and its dynamic parameters. A 99% Pearson coefficient was obtained between the perceptual and objective evaluation. The mean square error between the confusion matrices of both experiments is less than 5%.

Video material is being carried out using the web interface and equivalent objective experiment to near speech is being performed with far field speech.

5. Phonetic and Prosodic Labeling

SEV has been phonetically labelled using HTK software (Gallardo-Antolin et al., 2007) in an automatic way. In addition to this, 5% of each sub-corpus in SEV has been manually labelled, providing reference data for studies on rhythm analysis or on the influence of the emotional state on automatic phonetic segmentation systems. EGG signal has also been automatically pitch-marked and, for intonation analysis, the same 5% of each sub-corpus has been manually revised too.

Video data has been aligned and linked to speech and text data, providing a fully-labelled multimedia database.

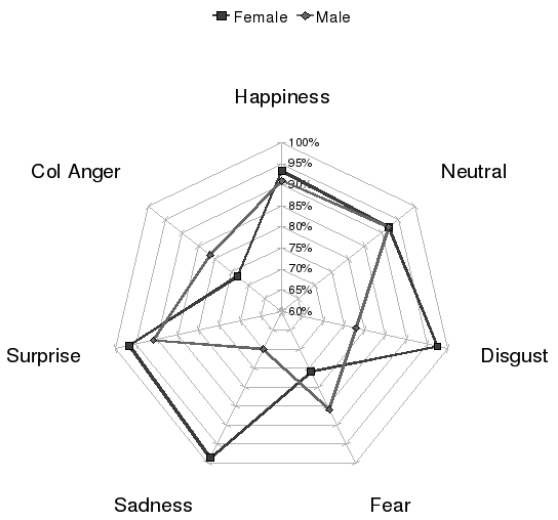


Figure 5: Average of perceptual identification rates for both speakers and emotion

6. Conclusions

In this paper we have presented SEV, a multimedia and multi-purpose database for research on emotional speech and video. Although part of the database was specifically designed for high-quality emotional speech synthesis and speech conversion, recorded data can be used for other emotion-related tasks such as emotional visual speech, close-talk or far-field emotion detection and emotional speech recognition or video-based emotion identification. Regarding speech, SEV covers a huge variety of contexts and situations, including recordings for Diphone-based or unit selection synthesis, and special sub-corpora for complex prosodic modelling.

Finally, the whole speech database has been evaluated through highly-correlated objective and automatic emotion identification tests. 89% of the recordings have been validated, achieving a general inter-labeller agreement higher than 0.95. 60% of the recordings were evaluated as intense or very intense in an emotional-intensity subjective test.

7. Acknowledgements

This work has been partially supported by the Spanish Ministry of Education & Science under contracts EDECAN (TIN2005-08660-C04-04) and ROBONAUTA (DPI2007-66846-c02-02).

8. References

R. Barra-Chicote, J.M. Montero, J. Macias-Guarasa, L.F. DHaro, R. San-Segundo, and R. Cordoba (2006). *"Prosodic and segmental rubrics in emotion identification"*. In Proceedingsof ICASSP, pages 1085–1088.

R. Barra-Chicote, J.M. Montero, J. Macias-Garasa, J. Gutirrez-Arriola, J. Ferreiros, and M. Pardo (2007).*"On the limitation of voice conversion techniques in emotion identification"* In Proc. of Interspeech.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss (2005). *"A Database of German Emotional Speech*". In Proc. of Interspeech.

G. Castellano, L. Kessous, and G. Caridakis (2007). *"Multimodal emotion recognition from expressive faces, body gestures and speech"*. In Humaine. International Conference on Affective Computing and Intelligent Interaction.

L. Chen et al.(2005). *"VACE Meeting Corpus"*. Lecture Notes in Computer Science. Pages 40-51.

N. Mana et al. (2007). *"Multimodal corpus of multi-party meetings for automatic social behaviour analysis and personality traits detection"*. In ICMI '07 Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, pages p. 9–14.

A. Gallardo-Antolin, R. Barra-Chicote, M. Schrder, S. Krstulovic, and J.M. Montero (2007)."In *Automatic Phonetic Segmentation f Spanish Emotional Speech"* In Proc. of Interspeech.

J.M. Montero, J. Gutirrez-Arriola, S. Palazuelos, E. Enrquez, and J.M. Pardo (1998). *"Spanish emotional speech from database to tts"*. In Proceedings of ICSLP, pages 923–925, September.

J.M. Montero, J. Gutierrez-Arriola, R. Cordoba, E. Enriquez, and J.M. Pardo (2002). *"The role of pitch and*

tempo in emotional speech". In Improvements in speech synthesis. Ed. Wiley and Sons, pages 246–251

Marc Schoder. (2004). *"Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis"*. Ph.D. thesis, Institut fur Phonetik. Universitat des Saarlandes, Saarbruken.

K.R. Schrerer and H. Ellgring (2007). *"Multimodal expression of emotion: Affect programs or componential appraisal patterns?"*. In Emotion. 2007 Feb ; vol 7 (1), pages. 158-171

N. Sebe, I. Cohen, and T.S. Huang (2005). *"Mutimodal emotion recognition: Handbook of pattern recognition and computer vision"*. In World Scientific.

SET	TEXT	*UTT/emo*	Length (*min/emo*)	<W>	<A>
LOGATOMOS	*Isolated words*	570	18	-	-
SESII-A	*Short sentences*	45	6	5	21
SESII-B	*Long sentences*	84	17	15	65
QUIJOTE	*Read speech*	100	22	16	70
PROSODIA 1	*A speech*	25	8	26	125
PROSODIA 2	*Interview (short anwsers)*	52	8	10	44
PROSODIA 3	*Interview (long anwsers)*	40	10	20	87
PROSODIA 4	*Question anwsering*	117	10	4	19
PROSODIA 5	*Short dialogs*	142	13	4	22
TOTAL		*1175*	*112*		

Table 1: Features related to SEV size

IrcamCorpusExpressivity: Nonverbal Words and Restructurings

Grégory Beller, Christophe Veaux and Xavier Rodet

IRCAM
1. place Igor Stravinsky
75004 Paris, France
{beller, veaux, rodet}@ircam.fr

Résumé

In this paper, we present the various constituents of a spoken message which allow the observation of expressivity in speech. These constituents are joined into the perspective of the double coding of the speech, which distinguishes the linguistic channel of the paralinguistic channel in a spoken message. Among this last channel, several phenomena seem to participate in the demonstration of the expressivity: The prosody, naturally, but also the nonverbal sounds, as well as of possible restructurings. In a second part, we introduce the expressive French multi-speaker corpus: IrcamCorpusExpressivity. Several steps of labeling and analysis allow the examination of this corpus under the various angles corresponding to the constituents of the spoken message. These results can be used to improve the tasks of recognition, transformation and synthesis of the expressivity in the speech, and so contribute to the anthropomorphisation of the Human-machine interfaces.

1. Introduction

Human-machine interfaces based on voice processing allow to obtain good rates in neutral speech recognition and to supply an understandable quality with synthesis. The introduction of the treatment of the *expressivity* in these tasks bring the researchers today to analyze the *expressive* speech (Bulut et al., 2007) (Yamagishi et al., 2005). The term *expressivity* is, here, defines as a level of information in the communication (Beller, 2008b). This level groups together the external demonstrations, simulated or not, which are attributable to internal states. Among these internal states are included the emotions, the attitudes, the feelings, the humors as well as the other styles which compose the range of actor's performance style. Our goals is to transform the expressivity of a neutral given utterance (recorded or synthesized) (Beller and Rodet, 2007) and are intended for artistic purposes (musical composition, contemporary theater, dubbing of cinema, animation, avatars and robots).

To do it, studies on the prosodic variations that can be attributed to the changes of expressivity were led. In particular, the introduction of new paradigms of analysis allowed to estimate the influence of the expressivity on the speech rate (Beller et al., 2006) and on the degree of articulation (Beller et al., 2008a). These studies notably showed the importance of the breaths which are part of nonverbal sounds. So other phenomena as purely prosodic participate in the communication of the expressive information. Some of these phenomena are emphasized here, notably thanks to the observation of an expressive French multi-speaker corpus : IrcamCorpusExpressivity.

After a theoretical introduction of the various constituents of the speech, we describe in a exhaustive way the corpus IrcamCorpusExpressivity realized within the VIVOS[1] project. The various levels of manual labeling are detailed to supply dictionaries created in this occasion and with the aim of showing certain tendencies according to the expressivity. The study of the continuous parameters of the prosody, which is one of our major subjects usually, is voluntarily put aside in this paper, so as to leave more place with the presentation of the corpus, as well as on the examination of the various levels of labeling relative to the other constituents of the speech.

[1]VIVOS :`http://www.vivos.fr`

2. Constituents of the speech

To observe the influence of the expressivity on the speech, we describe in this part various phenomena which the verbal communication implies and called constituents of the speech. They are summed up in the figure 1 and connected together with the perspective of the double coding of the speech, proposed by Fónagy (Fónagy, 1983). This perspective differentiates the linguistic channel of the paralinguistic channel. The linguistic channel is carrier of the semantic information and can be transcoded, without loss of information, in a text. The paralinguistic channel vehicles other levels of information that those carried by the linguistic channel, as the speaker identity, the speaking style, the modality, the prominence and, indeed on, the expressivity (Beller, 2008b).

2.1. Linguistic channel : verbal words and syntax

The linguistic channel brings the verbal words and their syntactical relationships. From an acoustic point of view, it is supported by sequences of segments, called *phones*. These phones are realizations of phonemes, which constitute the symbolic closed dictionary of differentiable sounds of a language. A verbal word possesses a meaning and a linguistic transcription. It can be written by use of a term stemming from the dictionary of common and proper nouns of a language. It thus depends on the sociocultural standards, quite as the syntax which depends on the grammar and which is also a part of the linguistic channel.

2.2. Paralinguistic channel

Among the paralinguistic channel, we discern the prosody, the nonverbal words and certain restructurings. All these elements are carriers of information others than linguistic.

2.2.1. Nonverbal words

The nonverbal words are sounds deprived of linguistic functions. By opposition to the verbal word, a nonverbal word does not possess usual transcription. However, it is not rare to find phonetic-spelling transcriptions of these sounds as ”ah ah ah” or ”laughter” to describe the presence of a laughter in a text (from the comic-strip to the novel, by way of the script of a play). It is because of this semantic dimension relative to the expressivity that we speak here of nonverbal words and not nonverbal sounds.

As the nonverbal words do not possess standardized transcription, they are with difficulty describable otherwise than by reproduction. In spite of a big variety, we distinguish among the nonverbal words, ”fillers” (laughter, scream, tear...), the breaths (inspirations, stops, expirations...) (Beller et al., 2006) and the other noises (guttural, nasal, of mouth...). It seems that these nonverbal words are of rich meaningful for the expressivity (Schroeder et al., 2006). The sadness can be only perceived by a tear and the fear, only by a scream, without the support of any verbal word. More finely, an informal perceptive experiment shows that the simple local addition of a breath in the middle of a neutral sentence, can change the perceived expressivity of the whole utterance[2]. The expressive power of the nonverbal words is such, that speech synthesizers begin to generate them (Beller, 2008a), so as to increase the naturalness and the expressivity of the synthesis. It requires, among others, the definition of standard for their transcriptions. The recent attempts base for the greater part on extensions of the SSML[3] language (Eide et al., 2004) (Blankinship and Beckwith, 2001).

2.2.2. Restructurings

The way are temporarily ordered the verbal and nonverbal words is informative. This is well known in the case of the verbal words temporal organization of which is defined by syntactical constraints. In the case of a spontaneous communication, these words can however not respect any more the order governed by the grammatical rules while preserving them syntactical functions. Indeed, the contiguity between nonverbal and verbal sounds force these last ones to possible temporal reorganizations called *restructurings*. So, although the syntax adjacent to the linguistic message organizes a priori the words and thus the sequences of phones, numerous not grammatical restructurings come into play, as the repetition of phones, syllables, whole word either even whole propositions (resetting). In spontaneous speech, the repetition which is frequent does not affect necessarily the understanding of the words and their syntactical relations. On the other hand, it can be a demonstrator of the hesitance or the confusion which are categories of expressivity. Other restructurings are carriers of sense for expressivity, while they are generally considered as *disfluencies* for the neutral speech (Piu and Bove, 2007) and concern the pronunciation : The *coarticulation*, the caesura, the connection and the elision are examples.

[2]Examples listenable to at : `http://www.ircam.fr/anasyn/beller`

[3]SSML : Speech Synthesis Markup Language : `http://www.w3.org/TR/`

2.2.3. Prosody

The stream of speech is thus a sequence of verbal and nonverbal words all organized by the conjugate action of the syntactical rules and the possible restructurings. At the same time, the acoustic realization of all these sound segments is ”modulated” by the prosody. If this is well known as regards the verbal words, it remains true for the nonverbal words as the laughter, for example (Beller, 2008a). The prosody includes suprasegmental phonological features the temporal span of which exceeds the boundaries of the phone (the syllable, the accentual group, the word, the clitic, the breath group, the prosodic group, the sentence...) and which do not annul the comprehensibility (that is that they do not deprive a phone of its membership in a phonetic category). Five characteristic features are generally quoted in the literature as the five dimensions of the prosody (Pfitzinger, 2006) :

– intonation : fundamental frequency, pitch
– intensity : energy, volume
– speech rate : flow, rhythm, speed of delivery
– degree of articulation : pronunciation, configurations of the vocal tract, dynamics of formants
– phonation : glottis signal, voice quality (pressed, normal, breathy), vibratory mode (fry, normal, falsetto), voicing frequency...

For a half a century of study of the neutral speech, the prosody was often reduced to the intonation. The intonation so benefited from a lot of attention and modelling, because, easy to observe, it allowed it only, bringing to the foreground functions of the prosody (modality, emphasis). The case of the expressive speech seems to require more strongly the observation of the other dimensions (Campbell and P.Mokhtari, 2003). Finally, of part its continuous character in the time, the prosody accompanies the production of verbal and nonverbal sounds and also interacts with the syntax and the restructurings.

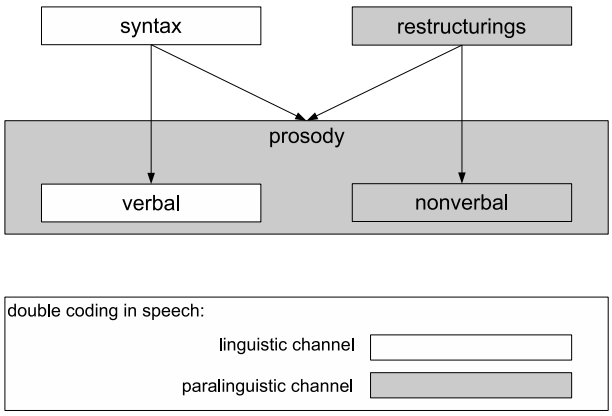


FIG. 1: Representations of the constituents of the speech. The verbal words and the syntax establish the linguistic channel. The nonverbal words, the prosody and the restructurings are the vectors of the paralinguistic channel.

A carrier vocal message of several levels of information is established by a sequence of verbal sounds and nonverbal sounds, all modulated by the prosody and organized by the conjugate action of the syntax and the restructurings (see fi-

gure 1). These paralinguistic phenomena are especially observable in the spontaneous speech, even more in the spontaneous dialogue and of advantage still in the case of the expressive speech as shows it the examination of the expressive corpus IrcamCorpusExpressivity.

3. IrcamCorpusExpressivity

The corpus IrcamCorpusExpressivity consists of recordings of four actors : Jacques, male, storyteller/comedian (∼40 years), Philippe, male, comedian dubber (∼40 years), Olivia, female, comedian dubber (∼25 years) and Danielle, female, comedian dubber (∼50 years). Every recording was guided thanks to a computing interface allowing a simplification of the recording process. This interface possesses a screen presenting the sentence, the expressivity and the intensity to be realized. The comedian starts and ends the recording thanks to a pedal. This interface also facilitates the post-production because it allows the synchronization, the labeling and the segmentation of the corpus as this one is recorded. So the actor can make a mistake or begin again without that it entails gaps. The comedians were recorded in the same conditions and in the environment which they know because it is their workroom. The studio of dubbing presents the advantage of an appropriate acoustics, being enough reverberating. So the actors feel less vocal fatigue than in an anechoic chamber, which possesses an unusual and particularly dry acoustics. A static microphone wearing an anti-pop filter allowed the acquisition of the data in ADAT[4] quality. Data stemming from an Electro-Glotto-Graph (EGG) are also available on certain parts of the corpus. In the end, more than 500 utterances were taken in by actor, forming a corpus of total duration about 12 hours of expressive speech.

3.1. Recited Text

The recited text were extracted from a French corpus of twenty sets of ten sentences. Every set is phonetically balanced (Combescure, 1981) :

1. C'est un soldat à cheveux gris.
2. Alfred pris la tête de l'expédition.
3. Il ne pourra pas me voir si j'éteins la lampe.
4. Il entre avec sa chandelle, dans la vielle chambre.
5. Le nez du personnage s'abaisse, au-dessus de sa moustache.
6. Vous êtes vraiment obéissant !
7. En attendant, c'est moi qui vais ouvrir.
8. Je ne pourrai jamais, me plier à son autorité.
9. Tout le monde sait que tu es la meilleure.
10. Je me demande, où se trouve cet endroit ?

This set was especially chosen because it contains neutral sentences with regard to the expressivity. That is that these sentences make sense, with every the expressivity with which they are pronounced. The prominence of some syllables was indicated to the actors by the punctuation and by the usage of uppercase characters. It allowed to vary the places of prominence and thus the resultant prosody, and to "congeal" the accentuation of the sentence to fix the semantic contents from a repetition to the other one.

3.2. Expressivity

The range of the wanted expressive categories was defined at the starting point of the VIVOS project, taking account the needs of a dubbing studio, of a commercial TTS synthesizer company and of an embedded video games company :

– Neutral
– *introvert anger* : contained or cold anger
– *extrovert anger* : explosive or warm anger
– *introvert happiness* : sweet or maternal happiness
– *extrovert happiness* : explosive or enthusiastic happiness
– *introvert fear* : contained or tetanic fear
– *extrovert fear* : explosive or alarming fear
– *introvert sadness* : contained sadness
– *extrovert sadness* : explosive or tearful sadness
– discretion
– disgust
– confusion
– positive surprise : the speaker is pleasantly surprised
– negative surprise : the speaker is unpleasantly surprised
– excitement

So as to be able to represent the recorded expressivities in a dimensional space axes of which are the valence (positive vs negative), the intensity (degree of intensity of the expressivity) and the activation (introversion vs extraversion) (Schroeder, 2003), we asked the actors to express the primary emotions (Ekman, 1999) with several degrees of intensity and according to two versions relative to the introversion and to the extraversion. For the last expressivities (in normal character), the comedians directly said all the text with the level of intensity the strongest possible. For the expressivities in italics, the degree of intensity was varied according to five levels. The progress of the recording is described by the following procedure. For a given expressivity, the speaker utters the first sentence in a neutral way. Then she/he repeats five times this sentence, with the wished expressivity, by increasing her/his degree of intensity. Then she/he moves to the following sentence and begins again this progress. Finally, she/he repeats this plan with the other expressivities. This procedure notably allows to obtain an intensification of the expressivity without that the speaker is to read again the text every time. From an intensity to the other one, neither the sentence, nor its accentuation changes, letting seem only the variations attributable in the intensity of the expressivity. The actors had for explicit order not to vary the pronunciation of their realizations corresponding to a sentence. This so as to minimize the variations due to the phenomena of restructuring which complicate the comparison of an expressive utterance with its neutral version and the building of conversion prosodic model (Tao et al., 2006) (Hsia et al., 2007). Interestingly and as it will be shown further, some restructurings appear despite this order. Nonverbal sounds were recorded then separately at the end. The following fillers was collected : "ah", "oh", "laughter", "tear", "fear", "panic", "enjoyment", "euh", "interrogation", "argh", "effort", "running", "hhh", "fff", with several realizations according to the ex-

[4]ADAT : Alesis Digital Audio Tape : 16bit, 48KHz

pressivity, for some of them.

3.3. Collected data

The data collected during the recording consist of audio files for every sentence and corresponding XML[5] files, containing the labeling of the expressivity (category and intensity), of the recited text, and of the information relative to the identity of the speaker (age, sex, name). These starting data have been manually labeled : phonetic segmentation, paralinguistic labeling and prominence labeling. Then symbolic analyses derived from these labels and, finally, acoustic analyses of the prosody have been processed. All the data which we are afterward going to describe, are stored and made accessible by IrcamCorpusTools, a database management system involving a powerful language of request (Beller et al., 2008b).

3.4. Phonetic segmentation

The phonetic segmentation of this corpus is, actually, a semi-phonetic segmentation (thus more precise). Indeed, a phone consists of two semiphones whose borders also allow to establish diphones (for the analysis and the synthesis). This segmentation was initialized by an automatic method (Lanchantin et al., 2008). This one leans on of multiple phonetizations of the text (Bechet, 2001) and implies a neutral French multi-speaker corpus (Lamel et al.,). This tool allows a fast and automatic segmentation which is not regrettably sufficient in the case of expressive speech. This *bootstrap* segmentation was thus manually checked and corrected by a phonetician. Then this correction was verified by another one. Not only the borders were moved but labels were also changed when it turned out necessary. The used code is the XSampa, which is an ASCII version of the IPA[6] chart.

Correcting the *bootstrap* segmentation, the phoneticians noticed numerous differences with the predictions of the machine (trained on neutral speech). For certain expressivities, expected phones was so differently realized that they were relabelized by other phonemes (opened /E/ moved to closed /œ/, for instance). Furthermore, some disappeared whereas other, unexpected, appeared. That is why, although all the expressivities were supported by the same text, we expected disparities in the posterior distributions of labeled phones, possibly attributable to the expressivity. However, all actors included, the average proportions of appearance of phones grouped together into phonological classes (see figure 2), do not show significant differences, function of the expressivity. Only confusion shows significant differences, cause by the numerous repetitions, as shown by the analysis of paralinguistic labels (see next section). On the other hand, direct local comparisons (and not statistical, as shown here) of the phonetic labels of the ideal phonemic sequence deduced from the text, and of the labeling of the realized phonetic sequence can allow to deepen this study.

3.5. Paralinguistic segmentation

Simultaneously in the operation of manual correction of the phonetic segmentation, a layer of supplementary labels was

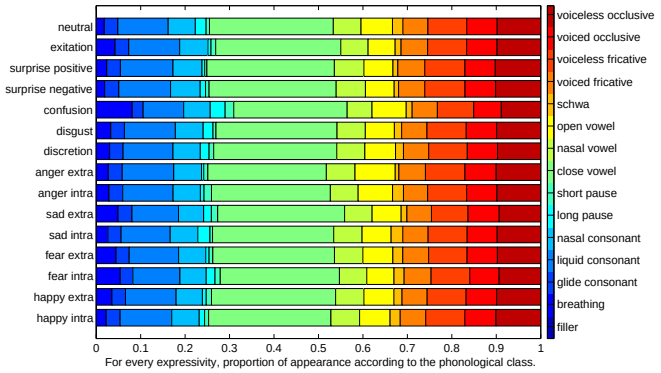


FIG. 2: All actors included. Average proportions of appearance of phonological classes by expressivity.

produced so as to supply paralinguistic information. This information notably describes the used nonverbal sounds, the possible restructurings and the diverse particular phonatory or prosodic phenomena. This stage of labeling required, once the stage of phonetic segmentation ended, a second pass to homogenize the labels. Indeed, because no dictionary of paralinguistic labels for this type of phenomena was defined a priori, the vocabulary employed by the annotators evolved according to the task and thus, from a corpus to the other one. The dictionary which we subject here, was thus the object of several inter-annotators discussions (and intra) and seems to gather the most important labels :

- Nonverbal sounds, breaths and voicing of the phonation :
  - [˚] : inspiration
  - [˚˚] : expiration
  - [nz] : nasal breath
  - [bx] : non vocal noise annoying signal analysis
  - [bb] : mouth noises
  - [ch] : whisper, instability of the voicing during the phonation, partial devoicing
  - [nv] : not voiced : total absence of vocal folds vibration
  - [ph] : transition : label indicating a nonverbal zone in continuity with a verbal zone (often short, but crucial for expressivity). In most of the cases, we meet this phenomenon either just before the first semiphone of a breath group, or just after the last semiphone of a breath group
- Pitch and guttural effects :
  - [fp] : pitch effects : sudden pitch variation often upward, concerning mostly only one semiphone, sudden change of vibratory mode ”normal” ↔ ”falsetto”
  - [fg] : guttural effect : audible glottal stops or starts, cutting of glottis excitement, sudden change of vibratory mode ”normal” ↔ ”fry”
  - [fi] : other effects than guttural or of pitch
  - [nt] : not transcribable : label put compared to the phonetic segmentation allowing to mean the doubt as for the attribution of the phone in a phonemic category
- Restructurings :
  - [lg] : length : phone abnormally long
  - [cu] : caesura : label applied to a silent phone by place (jerky phonation)

[5] XML : eXtensible Markup Language

[6] IPA : International Phonetic Alphabet

– [rp] : repetition : indexation of repeated phone or group of phones (up to 9 repetitions have been observed in the corpus : rp1, rp2, ..., rp9)

Finally, these various labels are composables thanks to the usage of the symbol [/] who allows to elaborate complex pattern from the given basic labels. For example, a voiced inspiration in falsetto mode starting a vowel is annotated by [˚/fp/ch] (seen in the extrovert fear case, for example), whereas a nasal expiration will be represented by [˚˚/nz]. Finally the interaction with the layer containing the phonetic segmentation is strong, because the same label put compared to a silence or to a phone will not mean the same thing.

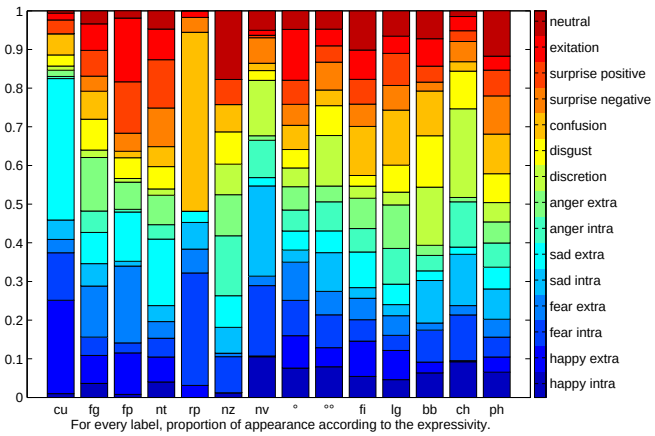


FIG. 3: All actors included. Average proportions of used paralinguistic labels according to the expressivity.

The figure 3 presents for each of the paralinguistic labels, the proportion of every expressivity. The more an expressivity contains a label in a recurring way (with regard to the others), the more the height of its associated rectangle is big. So, we observe that the caesura [cu] was strongly employed on the labeling of utterances expressed with extrovert sadness (jerky phonation). This expressivity contains so a lot of pitch effects [fp] and of no transcribable phones [nt]. In fact, several sentences are almost unintelligible because of a too weak degree of articulation (Beller et al., 2008a). The extrovert anger is marked by the presence of guttural effects [fg] as well as of nasal expirations [nz] (like the introvert anger and the disgust). Repetitions [rp] appear mainly for the introvert fear, and the confusion. This last expressivity also distances itself by numerous phonemes abnormally long [lg], like the ”angers” and the ”happinesses” (Beller et al., 2006). The discretion and the introvert sadness contain numerous markers of weak voicing ([nv], [ch] and [ph]). Furthermore, the inspirations [˚] are less labeled (perceived) compared to expirations [˚˚] in the case of the discretion. Finally the negative surprise seems less voiced that the positive surprise and presents less inspirations than expirations with regard to that last expressivity. Other numerous interpretations are possible and can, there also to be supported by more detailed local examinations.

3.6. Prominence labeling

From the phonetic segmentation, a rule based syllabifier (Veaux et al., 2008) produces a segmentation in syllables. The syllable plays a particular role in the prosody, notably because it is the smallest pronounceable prosodic group. Of a perceptive point of view, syllables distance themselves according to their levels of prominence. The prominence reflects an acoustic contrast (culminance, distinction, demarcation), performing several functions. First of all, it shows the accentuation of certain syllables which can be defined linguistically (Lacheret-Dujour and Beaugendre, 1999). The prominence plays sometimes also a role in the disambiguation of the sense (pragmatic accent). Finally, the prominence serves for emphasizing certain elements of the utterance (accent of focus, of emphasis, of insistence). A single annotator labeled the degree of prominence of the syllables of the whole corpus. The used scale consists of four levels :

– [UN] : indefinite or silence/pause (considered here as a syllable)
– [NA] : not prominent
– [AS] : secondary prominent
– [AI] : prominence with emphasis
– [AF] : final prominence (regular in French)

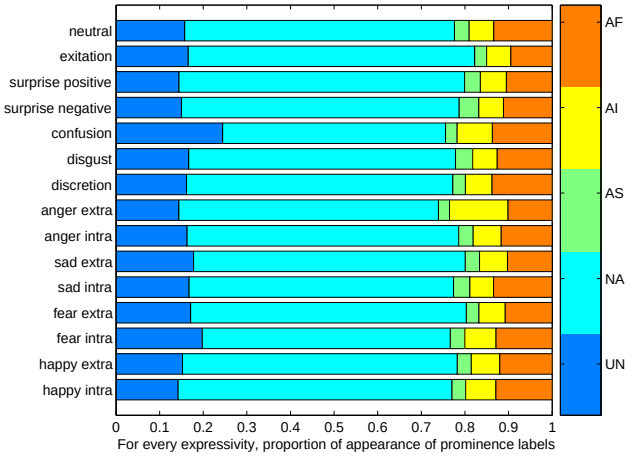


FIG. 4: All actors included. Average proportions of used prominence labels according to the expressivity.

A similar study to the previously presented ones concerning the distribution of these labels according to the expressivity does not show tremendous significant variations (see figure 4. Only the extrovert anger seems to distinguish itself from the other expressivities by a bigger proportion of [AI] labels. Indeed, syllables expressed with extrovert anger seems perceived more often as prominent. It can be explain, partially, by a hyperarticulation (Beller et al., 2008a) which provokes a detachment of consecutive syllables that become all prominent since they are all demarcated.

4. conclusion

In this paper, we presented, at first, a theoretical point of view allowing the observation of the expressivity in the speech. This point of view was included in the perspective of the double coding of the speech, which distinguishes the linguistic channel of the paralinguistic channel in a spoken message. Among this last channel, several phenomena seem to participate in the communication of the expressivity : the prosody, naturally, but also the nonverbal sounds,

as well as of possible restructurings. In a second part, we introduced the expressive French multi-speaker corpus : IrcamCorpusExpressivity. Several labelings and analyses allowed the examination of this corpus under the angle of the various phenomena belonging to the paralinguistic channel. Few differences attributable to the expressivity are visible in the phonological distributions, as well as in the distributions of the levels of prominence. However, these results are to be minimized because the actors had exactly for explicit order, not to make vary these constituents, but only the prosody. On the other hand, if they also had for order to avoid the usage of nonverbal sounds and restructurings, nevertheless this constituents appears frequently in the corpus. Their examination allowed to bring to light that certain expressivities distinguish themselves by strong apparences of some of these constituents. As if some of them required the use of nonverbal sounds and restructurings besides the prosodic variations to be expressed. To validate these findings, a similar study on corpus not basing on such orders is to be made. Nevertheless, these various results can be already used to improve the tasks of recognition, transformation and synthesis of the expressivity in the speech, and so, contribute to the anthropomorphisation of the Human-machine interfaces.

5. Acknowledgments

This work was partially funded by the French RIAM network project VIVOS. The authors wish to thank the actors involved in this study for their performances, as well as the various annotators having participated in the labeling of the data.

6. References

Frederic Bechet. 2001. Liaphon : un système complet de phonétisation de textes. In *Traitement Automatique des Langues - TAL*, number 1, pages 47–67.

Grégory Beller and Xavier Rodet. 2007. Content-based transformation of the expressivity in speech. In *ICPhS*, Saarbr ucken, August.

Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet. 2006. Speech rates in french expressive speech. In *Speech Prosody*, Dresden, may. SproSig, ISCA.

Grégory Beller, Nicolas Obin, and Xavier Rodet. 2008a. Articulation degree as a prosodic dimension of expressive speech. In *Speech Prosody 2008*, Campinas, May.

Grégory Beller, Christophe Veaux, Gilles Degottex, Nicolas Obin, Pierre Lanchantin, and Xavier Rodet. 2008b. Ircam corpus tools : Système de gestion de corpus de parole. *TAL, to appear*.

Grégory Beller. 2008a. Semi-parametric synthesis of speaker-like laughter. *to appear*.

Grégory Beller. 2008b. Transformation of expressivity in speech. In Peter Lang, editor, *The Role of Prosody in the Expression of Emotions in English and in French*. Peter Lang.

Erik Blankinship and Richard Beckwith, 2001. *UIST '01 : Proceedings of the 14th annual ACM symposium on User interface software and technology*, chapter Tools for expressive text-to-speech markup, pages 159–160. ACM, New York, NY, USA.

Murtaza Bulut, Sungbok Lee, and Shrikanth Narayanan. 2007. a statistical approach for modeling prosody features using pos tags for emotional speech synthesis. In *ICASSP*.

N. Campbell and P.Mokhtari. 2003. Voice quality : the 4th prosodic dimension. In *XVth Int. Congress of Phonetic Sciences*, volume 3, pages 2417–2420, Barcelona.

Pierre Combescure. 1981. 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 56 :34–38.

E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli. 2004. A corpus-based approach to <ahem/> expressive speech synthesis. In *5th ISCA Speech Synthesis Workshop*.

P Ekman, 1999. *The Handbook of Cognition and Emotion*, chapter Basic Emotions. John Wiley & Sons, Ltd.

I. Fónagy. 1983. *La vive voix : essais de psychophonétique*.

Chi-Chun Hsia, Chung-Hsien Wu, , and Jian-Qi Wu. 2007. conversion function clustering and selection for expressive voice conversion. In *ICASSP*.

A. Lacheret-Dujour and F. Beaugendre. 1999. *La prosodie du Francais*. CNRS langage.

L. Lamel, J. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french.

Pierre Lanchantin, Andrew C. Morris, Xavier Rodet, and Christophe Veaux. 2008. Automatic phoneme segmentation with relaxed textual constraints. In *Language Resources and Evaluation Conference (LREC2008)*, volume ND, Marrakech, Maroc, Mai.

H.R. Pfitzinger. 2006. Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction. In H Hoffmann, R. ; Mixdorff, editor, *Speech Prosody*, number 40 in Abstract Book, pages 6–9, Dresden.

Marie Piu and Rémi Bove. 2007. Annotation des disfluences dans les corpus oraux. In *RECITAL*.

M. Schroeder, D. Heylen, and I. Poggi. 2006. Perception of non-verbal emotional listener feedback. In *Speech Prosody 2006*, Dresden, Germany.

Marc Schroeder. 2003. *Speech and Emotion Research : An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. Ph.D. thesis, University of Saarland.

Jianhua Tao, Yongguo Kang, and Aijun Li. 2006. Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1145 – 1154, July.

Christophe Veaux, Grégory Beller, and Xavier Rodet. 2008. Ircamcorpustools : an extensible platform for speech corpora exploitation. In *LREC*, Marrakech, Maroc, Mai.

J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi. 2005. Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. In *IEICE Trans. on Inf. & Syst.*, volume E88-D, pages 503–509, March.

Dynamic Detection of Mood Propagation in Fan Groups

Gail L. Rein
Element K
Rochester, NY
E-mail: grein@rochester.rr.com

Dave Bruckmayr
Entrepreneur
Hoersching, Austria
db@bullfog.com

Paulo Barthelmess
Adapx
Seattle, WA
paulo@adapx.com

Abstract

This paper describes the corpus being developed to capture and analyze the mood of fan groups in a soccer stadium. Although our current focus is soccer fans, we are interested in mood propagation in groups of people engaged all varieties of human activity. Our goal is to understand the fundamental elements of mood change sufficiently that we can develop a machine-learning algorithm to dynamically and accurately detect and label these mood changes. In developing our crowd emotion corpus, we are finding that it is surprisingly difficult to identify the features that human beings subconsciously perceive, process, and use to label an emotional state. Cultural and activity-specific interpretations of these features are also critically important for a machine-learning algorithm.

1. The Research Problem

We are exploring mood propagation in groups of people. We want to know how and when mood reaches a "threshold" intensity—a point that once surpassed, the mood starts propagating. We are developing a method to detect and measure threshold events by their verbal and non-verbal triggers. In our inquiry, we are focusing on fan groups in soccer stadiums.

2. Why We Are Tackling the Problem

We are developing a "fan meter" that will show fan mood in real-time during a game. Threshold events provide the "mood data" that the meter will display. The corpus we are building will provide data for training macine-learning algorithms (Dietterich, 2003) at later phases.

A serious application for the fan meter is security; the meter could be used by security personnel to detect disturbances. Less serious, fun applications are to provide a form of biofeedback to fans in the stands and players on the field.

3. The Approach We Are Taking

We envision using the video and audio feeds that are available in stadiums as real-time input for the fan meter. In basic terms, we need to define three transformations:

1. Video data —> mood data
2. Audio data —> mood data
3. Mood data —> fan meter

3.1 Transformations 1 and 2

In a group of fans, there are agitators and responders. At a certain agitation level, moods start to propagate through the group, affecting the "moodless" members. In a preliminary analysis of footage of a soccer fan group (Bruckmayr, 2007), we were able to detect these propagation events at a distance without being able to see the faces of the individual group members. There are recognizable mixings of subconscious signals that precede mood propagation.

Labeling the emotional episodes is relatively easy and natural: the fans are happy, unhappy, excited, bored, etc. We are currently identifying the specific expressive features—such as shouting, clapping, booing—that define specific emotional episodes. We are also examining the correlation of these features with their associated emotional labels.

We have found that the crowd's behavior is harmonious and uniform or erratic when different groups within the crowd act differently. We refer to the former as an in-sync crowd and the latter as an out-of-sync crowd. Tables 1 and 2 below summarize some of the common expressive features and their corresponding emotional meanings in the context of the game for in-sync and out-of-sync crowds.

Expressive features	Emotional State	Comments
Low noise level	Attentive	Interested in what's happening on the field
Shouting, clapping, medium noise level	Responsive	Immediate reaction to happenings on the field
Chanting, cheer-leading, clapping, low noise level	Relaxed/ happy	Happens if the score is right
Chanting, jumping up and down, high noise level	Enthusiastic/ excited	The game is going better than expected

Table 1: Emotional features and their meanings when the crowd is in-sync.

Expressive features	Emotional State	Comments
Little movement in crowd, low noise level	Detached/ bored	Not interest-ed in what's happening on the field
Outbursts of booing, eruptive shouting, stretches of silence, medium noise level	Frustrated/ disappointed/ depressed	No hope for a turnaround, game seems lost
Shouting, one hand above head, two hands above head, high noise level	Agitated/ stressed	During deci-sion phases of the match
Booing, shouting, aggressive chanting, thumping on stand, whistling, very high noise level	Hostile/ freaked out/ unhappy/ upset/ angry	Fans feel robbed of goal or match
Silence interrupted by short stretches of chanting, low noise level	Shocked	Mostly after a goal against own team

Table 2: Emotional features and their meanings when the crowd is in out-of-sync.

Preliminary analysis reveals that mood can be associated with audio and visual manifestations. Mood can be correlated with the levels of fan-generated noise, the timing and pitch of shouts, screams, song, and applause. We are investigating questions such as, "Do agitators have to produce a certain noise level to induce mood propagation?" We are analyzing audience movement to recognize typical behavioral patterns that set the stage for mood propagation, such as stamping, throwing up arms, jumping, and thumping.

3.2 Transformation 3

Given the mood-segmented data, we plan to generate a summary representation using a compact graphical notation (Rein, 1991) that captures the socio/emotional, or affective, aspects of group interactions. This notation is an extension of Robert Bales' SYMLOG (Bales & Cohen, 1979) technique,[1] and was developed as a biofeedback display for group meetings. A notable attribute of this notation is that it describes emotions without placing value judgments on them; this is highly desirable so that emotions can be interpreted within the context of their real-time situations. One of the envisioned uses of this summarized data is to drive a visual display, or fan meter, that shows how fan mood shifts throughout a game.

To accomplish transformation 3, the emotional labels derived from transformations 2 and 3 will be mapped to SYMLOG space: downward/upward, negative/positive, and backward/forward. For example, shouting and clapping occurring together might be mapped as UPB since it is an emotionally expressive gesture (backward) that is dominant and powerful (upward) and friendly (positive). The number of fans who are shouting and clapping and the intensity of their shouting and clapping determines how strong the group is in each dimension.

3.3 Detection of Disturbances

If our approach is successful, we will be able to classify emotions that are "normal" or typical in soccer games. Detection of disturbances almost comes for free; possible disturbances will be emotional episodes that are atypical.

One of the authors is certified by the Union of European Football Associations (UEFA) as a security guard for the June 2008 European Soccer Championships in Austria and Switzerland. To become certified, he had to go through a rigorous training session. We are trying to get copies of the riot tapes that were used in this training. Extracting features from these tapes will allow us expand the training set to include disturbance examples.

4. Evaluation of Crowd Emotion Corpus Against the Workshop Questions

In this section, we evaluate our crowd emotion corpus against four of the questions examined in the workshop. Our intent here is to provide a useful characterization of the corpus against the issues, not to take the position that our approaches are appropriate for all corpora.

4.1 What are the appropriate sources?

The source of our corpus is clearly application-driven as it consists of emotions captured in real-time during soccer games. It is also naturalistic data, but narrow in scope as it represents emotions common at soccer games, not everyday life. This corpus is also culturally sensitive

[1] SYMLOG is an acronym that stands for SYstem for Multiple Level Observation of Groups. It is a theory and method for the analysis of small group interaction, based on four decades of research and testing by Robert F. Bales and his students. The early work is described in Bales' classic book, *Interaction Process Analysis* (1950). SYMLOG has been applied by both practitioners and other researchers to a wide range of applications (Polley et al., 1988), including self-analytic groups, classroom management, social work, group and family therapy, content analysis, international relations, attitude measurement, leadership, team building, and organizational development. The method required human observers who manually coded the interactions they observed, and this restricted the size of the groups that were observed. The theories, however, apply to human interactions in general. Using macine-learning algorithms to recognize and code the interactions makes it feasible to apply the method to very large social groups, such as fans in a stadium.

data. For example, when fans whistle in Austria, they are unhappy and upset; whereas when fans whistle in the U.S., they are happy and excited.

4.2 Which modalities should be considered, in which combinations?

Our corpus is created from two channels: video and audio. We are capturing expressive features such as shouting, chanting, clapping, jumping up and down, one hand above head, two hands above head, cheerleading, thumping the stands, booing, and whistling. At any point in time, each feature may be present in varying degrees: not present, some, medium, and lots. Thus, we consider the intensity of each feature as well as combinations of features. The mapping of these features to the SMLOG space of downward/upward, negative/positive, and backward/forward is culturally dependent.

4.3 What are the realistic constraints on recording quality?

Since we are interested in a group's emotion, not an individual's emotion, our corpus does not require high-quality recordings. We do not need, for example, to pick up on a raised eyebrow or a moan under one's breath. We are looking at the macro features of emotion rather than the micro features. On the other hand, recording may be challenging because large stadiums require multiple cameras for coverage, positioned strategically over a large area.

4.4 How can the emotional content of episodes be described within a corpus?

Emotional episodes are represented in our corpus at three levels:

1. The audio and visual features (e.g., shouting, clapping, jumping).
2. An associated emotion (e.g., bored, angry, happy).
3. A SYMLOG-based summary (downward/upward, negative/positive, and backward/forward).

We are using a scoring sheet that we are happy to share, to manually identify the emotional features (shouting, chanting, etc.) observed at each second of a recording. For each feature, we enter a number from 0 to 3, where 0 is not present, 1 is some, 2 is medium, and 3 is lots. This gives us a profile of the features that we then analyze for patterns, which in turn allows us to identify the emotional episodes. We then attach a descriptive label of the emotion for each of these episodes, and finally generate the SYMLOG-based classification.

5. Related Work

Others' research on group emotions includes multi-player games, encouraging more sociable behavior, and multimodal interfaces. Representative papers are cited here along with the types of insight they offer our research.

Emotional Flowers was a multi-player game in which players competed by using their positive emotions to grow flowers (Bernhaupt et al. 2007). The researchers observed that the emotions people used to control the game transcended the game and improved their everyday lives on an interpersonal level. This outcome is a form of mood propagation.

Foucault et al. (2007) designed and developed an agent-based system that collected information from office workers by asking seemingly benign questions. The collected information was then used to spread false, strange gossip with the desired outcome that this gossip would improve off-line sociability. This outcome is demonstration of indirect mood propagation. Gossip spreads information and opinions about it, and this in turn can induce mood.

Maynes-Aminzade et al. (2002) developed an interactive entertainment system that allowed members of a large audience to control an onscreen game of Pong by leaning left or right. The computer vision techniques developed by these researchers are of special interest to us.

There is also a significant body of literature related to game analysis, but it is focused on the activities of the players rather than the spectators (Assfalg et al., 2003; Baillie & Jose, 2004; Chang et al., 2001; Christmas et al., 2003; Kang et al. 2004; Leonardi et al., 2004; Liu et al., 2005; Masanori et al., 2007; Xiong et al., 2003).

6. Conclusions

Although our current focus is soccer fans, we are interested in mood propagation in groups of people engaged all varieties of human activity. Our goal is to understand the fundamental elements of mood change sufficiently that we can develop a machine-learning algorithm to dynamically and accurately detect and label these changes. To make practical progress on this goal, we felt it important to study intact, natural groups of people who are interacting in the context of some identifiable activity. Soccer fans was as good a context to start with as any.

In developing our crowd emotion corpus, we are finding that it is surprisingly difficult to identify the features that human beings subconsciously perceive, process, and use to label an emotional state. Cultural interpretations of these features are also critically important for a machine-learning algorithm.

There may well be a set of universal features that define basic human emotions. Comparing the corpora on a variety of human activities will identify such features. Hence, many of the questions posed for discussion at this workshop are valid because researchers working on a variety of corpora need to be able to compare their findings. From our work with soccer fans, we believe however that there will *always* be some features that are specific to an activity. As a community of researchers, we must not forget this in our enthusiasm to generalize and unify our findings.

7. Acknowledgements

The authors appreciate the thoughtful feedback and advice from the three anonymous reviewers.

This project is a true team effort, and all three authors contributed equally. Paulo Barthelmess, Ph.D. Computer Science, is an industrial HCI researcher and prototyper with expertise in intelligent adaptive systems and multimodal and perceptual systems. Dave Bruckmayr is a freelance journalist and idea developer with a special interest in mobile computing; he's also a soccer fan. Gail L. Rein, Ph.D. Information Systems, is a systems designer with expertise in multi-user interfaces, visual languages, and group and organizational processes.

8. References

Assfalg, J., M. Bertini, C. Colombo, A.D. Bimbo, and W. Nunziati (2003). Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding*, Vol. 92, Issues 2-3, pp. 285-305.

Baillie, M. and J.M. Jose (2004). An audio-based sports video segmentation and event detection algorithm. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop*, p. 110.

Bales, R.F. (1950). *Interaction Process Analysis: A Method for the Study of Small Groups*. University of Chicago Press, Chicago, IL, 1950 (reprinted 1976).

Bales, R.F., and S. Cohen (1979). *SYMLOG, A System for the Multiple Level Observation of Groups*. New York: The Free Press, Macmillian Publishing Co., Inc.

Bernhaupt, R., A. Boldt, T. Miriacher, D. Wilfinger, and M. Tscheligi (2007). Using emotion in games: emotional flowers. In *Proceedings of the ACE 2007 Conference*, June 13-15, Salzburg, Austria, pp. 41-48.

Bruckmayr, D. (2007). Footage of soccer fans at a club game in Hoersching Austria on 8/25/2007. Quicktime movies are archived at http://www.bullfog.com/soccer/soccerfans2.mov and http://www.bullfog.com/soccer/soccerfans3.mov.

Chang, S.F., D. Zhong, and R. Kumar (2001). Real-time content-based adaptive streaming of sports videos. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 139-146.

Christmas, W., E. Jaser, K. Messer, and J. Kittler (2003). A multimedia system architecture for automatic annotation of sports videos. *Computer Vision Systems* (Series: *Lecture Notes in Computer Science*, Vol. 2626/2003). Springer Berlin / Heidelberg.

Dietterich, T. (2003). Machine learning. In *Nature Encyclopedia of Cognitive Science*. London: Macmillan.

Foucault, B. P. Sengers, H. Mentis, and D. Welles (2007). Provoking sociability. In *Proceedings of the CHI 2007 Conference*, pp. 1557-1560.

Kang, Y.L., J.H. Lim, Q. Tian; M.S. Kankanhalli, and C.S. Xu (2004). Visual keywords labeling in soccer video. In *Proceedings of the 17th International Conference on Pattern Recognition* (ICPR 2004), Vol. 3, pp. 850-853.

Leonardi, R., P. Migliorati, and M. Prandini (2004). Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14, Issue 5, pp. 634-643.

Liu, S., M. Xu, H. Yi, L.T. Chia, and D. Rajan (2005). Multimodal semantic analysis and annotation for basketball video. *EURASIP Journal on Applied Signal Processing*, Vol. 2006, Article ID 32135, 13 pages.

Masanori, S. I. Yamada, H. Sumiyoshi, and N. Yagi (2007). Automatic real-time selection and annotation of highlight scenes in televised soccer. *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 1, pp. 224-232.

Maynes-Aminzade, D., R. Pausch, and S. Seitz (2002). Techniques for interactive audience participation. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces* (ICMI'02), 6 pages.

Polley, R.B., A.P. Hare, and P.J. Stone (Editors) (1988). *The SYMLOG Practitioner: Applications of Small Group Research*. Praeger Publishers, New York, NY.

Rein, G.L. (1991). A group mood meter. In *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*, Vol. 4, pp. 308-323.

Xiong, Z., R. Radhakrishnan, and A. Divakaran (2003). Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In *Proceedings of the 2003 International Conference on Image Processing* (ICIP), Vol. 1, pp. 5-8.

Ambiguous classification of emotional speech

Zhongzhe Xiao[1], Emmanuel Dellandrea[1], Weibei Dou[2] and Liming Chen[1]
[1] Laboratoire d'InfoRmatique en Images et Systemes d'information (LIRIS),
Département MI, Ecole centrale de Lyon,
36 avenue Guy de Collongue, 69134 Ecully Cedex, France;
E-mail: {zhongzhe.xiao; emmanuel.dellandrea; liming.chen}@ec-lyon.fr
[2] Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P.R.China;
E-mail: douwb@mail.tsinghua.edu.cn

Abstract

Speech emotion is high semantic information and its automatic analysis may have many applications such as smart human-computer interactions or multimedia indexing. Among the main difficulties encountered when developing an automatic recognition system are the definition of emotions and their subjective nature. Thus, with the consideration of a limited number of types of emotions, a certain emotional state can be between some pre-defined emotional states, while a too large number of types can lead to an insolvable classification problem. In this case, a classification with the management of ambiguous emotions is necessary. Such an approach is proposed in this paper: in order to make the automatic recognition of emotions as close as possible to the judgments produced by humans, we developed an ambiguous classifier which allows to give multiple labels to emotional speech. This approach has been evaluated on Berlin dataset and compared with multiple human judgments used as ground truth.
Keywords: speech emotion, ambiguous classification, evidence theory.

1. Introduction

The interest in expressive speech can be traced back to the early Greek and Roman manuals on rhetoric which were the basis of the later theory of emotional appeal in western philosophy (Scherer, 2002). In the 19th century, a new interest in the expression of emotion in face and voice was motivated by the emergence of modern evolutionary biology, particularly due to Darwin's research in 1872 on how animals and humans express and signal to others their emotions. Systematic research on the emotions started in the 1960s when psychiatrists renewed their interest in diagnosing affective states via vocal expression. Emotion psychologists, linguists, phoneticians, engineers and phoneticians also took part into the research of audio emotions from various aspects from the 1970s and then, the automatic detection of the emotions began to come into interest in the last few years.

Although many research works dealing with the notion of emotion have been made, there is still no universal agreement on the basic definition of emotions. The two traditional theories on emotions are the discrete and the dimensional emotion theories. For the discrete emotion theory, different numbers and different types of emotions are proposed by researchers. The term "big six" gained attention implying the existence of a fundamental set of six basic emotions while there does not seem to be any agreement on which six these should be (Scherer, 2002). When the terms of the emotions are applied to the music, they should be modified to fit the longer lasting emotion states as moods. In the dimensional emotion approach, different emotional states are mapped into a two or three-dimensional space (Pereira, 2000; Scherer, 2000; Scherer, 2002; Thayer, 1989). The two major dimensions consist in the valence dimension (or appraisal dimension, pleasant – unpleasant) and the activity dimension (or arousal dimension, or energy dimension, active – passive). A joint description of the emotion definition combining the two traditional theories is proposed in our work in which the discrete emotion states are distributed in a dimensional space.

In the scope of emotion recognition in audio signals, the problems following the emotion taxonomy elaboration concern the selection of acoustic features presenting the emotion or mood aspects and the classification algorithms. Further to the effective features commonly used in speech recognition and other work on music analysis it is necessary for the recognition task to find new features which have the ability to represent the emotional characteristics.

Moreover, as the emotions are subjective judgments of human beings, the border between the difference emotions are usually ambiguous. With the consideration of a limited number of types of emotions, a certain emotional state can be between some pre-defined emotional states, while a too large number of types can lead to an insolvable classification problem. In this case, a classification with the management of ambiguous emotions is necessary. Thus, we have developed a method based on the evidence theory, which performs the combination of the sub-problems classifiers. The possibilities of each of the emotional states are given as the recognition result.

The subjective nature of emotions can lead to emotional states contained in a certain segment of speech not being definite as one definite emotional state. Illustrated by dimensional emotion, basically, there are two problems to be highlighted: first, an emotion state is quite fuzzy and should be continuous, e.g. explosive happiness vs. calm happiness, or hot fury vs. cold anger; second, the judgment from the human being may be multiple as the feeling of emotion is subjective. In this paper, we tend to make a preliminary approach on the automatic recognition of multiple judgments to the emotions.

In order to justify the possibility of multiple judgments by human, we made human test on the Berlin dataset (Sendlmeier) (see section 3 for an introduction on Berlin dataset). Five human subjects were asked to classify the speech segments from the Berlin dataset according to the expressed emotion. Subjects were asked to mark emotional labels to the speech segments, in the case of difficult in judging the emotional state, two or three labels are allowed to be marked on the same utterance.

The average confusion matrices in human testing are listed in Table. 1. Since the utterances are allowed to be labeled with multiple emotions, the sums of each row in the confusion matrix are possible to exceed 100%.

According to the results in human testing, the emotion sadness is almost perfectly recognized. Utterances from every emotion have chances to be judged as neutral, especially for the emotions happiness and boredom. Utterances from anger and happiness have relatively higher chance to be confused between each other, and boredom tends to be misjudged as sadness.

In order to make the automatic recognition of emotions as close as possible to the judgments produced by humans, we developed an ambiguous classifier which allow multiple labels to emotional speech, and result from the multiple human judgments on Berlin dataset is used as ground truth with this approach.

In our work, different sub-classifiers concerning all the emotional classes or some of the classes are evaluated and the best a few classifiers are selected to form the ambiguous classifier based on the evidence theory of Dempster-Shafer (Shafer, 1990; Shafer, 1992; Fioretta, 2004). This theory is chosen for its applications for modeling and quantifying the assigned belief to facts by giving an order of confidence to these facts (Telmoudi, 2004), and the beliefs can be combined by orthogonal sum with the Dempster's combination rule. The way sub-classifiers are built is presented in (Xiao, 2007). The principle is inspired from the wrapper feature selection method SFS (Pudil, 1994) since it relies on the simple principle to add incrementally most relevant features. Its originality concerns the use of mass functions from the evidence theory which allows to merge elegantly the information carried by features, in an embedded way, and so leading to a lower computational cost than original SFS.

The automatic ambiguous classifier we propose here is only a preliminary attempt on the ambiguous recognition of emotions, only the possibility with the multiple judgments is considered in our experiments, while the continuity of the emotions is not yet investigated, and should be discussed in our future work.

The reminder of this paper is organized as follows. Section 2 presents the ambiguous classification scheme we have developed. Experimental results are presented in section 3. Finally, conclusions and perspectives are drawn in section 4.

	Human judge / Original label	Anger	Happiness	Fear	Neutral	Sadness	Boredom
Female	Anger	74.63	14.90	0	4.48	14.90	8.06
	Happiness	12.50	92.50	0	5.00	5.00	15.00
	Fear	10.34	0	96.55	10.34	13.79	13.79
	Neutral	0	0	2.50	95.00	10.00	20.00
	Sadness	0	0	0	2.86	100	8.57
	Boredom	0	0	0	22.22	2.22	93.33
Male	Anger	63.33	16.70	0	1.67	8.35	12.00
	Happiness	8.57	87.50	4.17	20.83	0	8.33
	Fear	7.70	0	96.15	3.85	0	0
	Neutral	0	0	0	100	0	0
	Sadness	0	0	0	5.88	100	0
	Boredom	0	0	0	25.40	17.65	67.65

Table. 1 Confusion matrix in human judgment for multi-possibility on Berlin dataset (%)

2. Ambiguous classification

The ambiguous classifier we propose relies on several sub-classifiers each of them dealing with a simple classification problems with fewer classes. Instead of placing the sub-classifiers hierarchically as in our previous approach (Xiao, 2007), the sub-classifiers process the classification in parallel, and the results from the sub-classifiers are combined with the fusion applying the Dempster's combination rule into belief masses of each of the emotional states. The aim of this approach is to make possible the multiple judgments for emotions in case of utterances with ambiguous emotions. That is to say, when the emotion states contained in a certain utterance is between some emotional states considered in the problem, all the emotional states with relatively high beliefs can be presented in the judgment.

The generation process of the ambiguous classifier is shown in Fig. 1. The N discrete emotional/mood states concerned in the classification are first assigned as a frame of discernment $\Omega=\{E_1, E_2, \ldots, E_N\}$. For example, for the emotion classification on Berlin dataset, Ω_{Berlin}={Anger, Happiness, Fear, Neutral, Sadness, Boredom}.

Three fundamental steps are taken in the generation of the ambiguous classifier. First, the possible sub-classifiers are proposed, then the sub-classifiers are evaluated and several best sub-classifiers are chosen according to the classification performance, the selected sub-classifiers then pass a fusion process to get the final beliefs of each emotions. The detailed processes in these three steps are listed as follows:

1) Step 1: building of possible sub-classifiers

The Ω is first divided into different groups of nonempty subsets, and each group can correspond to a sub-classifier. Suppose the n^{th} group of subsets G_n contains N_n subsets $A_{n_1},\ldots, A_{n_{N_n}}$, where the group G_n correspond to the n^{th} sub-classifier, and each subset A_{n_i} presents a class in the n^{th} sub-classifier. Each group satisfies:

$$G_n = A_{n_1} \cup ... \cup A_{n_{N_n}}$$

$$G_n \subseteq \Omega$$

$$A_{n_i} \neq \phi, 1 < i < N_n$$

$$A_{n_i} \cap A_{n_j} = \phi, i \neq j$$

The union of the subsets in each group are not obliged to equal to the frame of discernment Ω, as the missing emotion in one group may be derived from the fusion of other sub-classifiers.

For example, we can have the following possible groups to be classified as sub-classifiers on Ω_{Berlin}:

$$G_{i_1} = \{Anger, Happiness\} \cup \{Fear, Neutral\} \cup \{Sadness, Boredom\}$$

$$G_{i_2} = \{Anger\} \cup \{Happiness, Fear, Neutral\}$$

$$G_{i_3} = \{Happiness, Fear\} \cup \{Neutral, Sadness\}$$

Taken the group G_{i1} as example, it corresponds to a classification problem concerning three classes: the first class as Anger & Happiness, the second class as Fear & Neutral, and the third class as Sadness & Boredom. A sub-classifier on these three classes can be built based on this group.

In order to avoid excessive computational load, not all the possible groups of subsets are evaluated. Since the classification problems with fewer classes tend to get better performance under the same situation, the number of subsets (classes) in the groups is limited to $N_n \leq 3$.

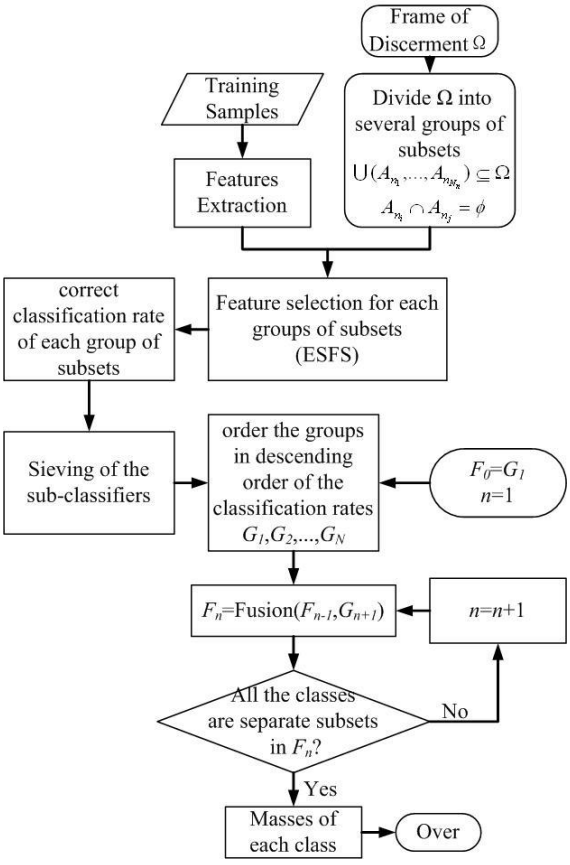


Fig. 1 Generation of the ambiguous classifier

The subsets in the same group are classified with the feature combination and selection scheme integrated in sub-classifiers proposed in (Xiao, 2007), which rely on the principle to add incrementally most relevant features by using mass functions from the evidence theory to merge the information carried by the features. For each group, the belief masses of each audio segment belonging to each of the subsets can be obtained with the selected combined feature.

Although the belief masses are given as the output of the sub-classifiers, the judgment of the class which the utterance belongs to is still needed to get the correct classification rates in order to make evaluations of the performance of the sub-classifiers. If the subset with the highest values is assigned as the judged class, correct classification rate R_{0n} can be calculated for the n^{th} groups of subsets G_{0n}.

2) Step 2: Sieving of the sub-classifiers

The groups of subsets are reordered in descending order according to the correct classification rates. From the rules of making groups of subsets (classes), up to several hundred groups of sub-classifiers will be evaluated for a classification problem with six or more classes in the previous step. As the final step will be fusions between the sub-classifiers according to the Dempster's rule of combination, a sieving step to the sub-classifiers is applied here before the fusions.

The Dempster's rule of combination, which is based on an orthogonal sum of the mass functions, requires the mass assignments to be independent to each other. Since the feature selection and classification of sub-classifiers on different groups of emotional subsets (classes) are processed separately, we assume the sub-classifiers based on the groups with distinct classes are approximately independent to each other, because the classes to be discriminated in these sub-classifiers are very different to each other, and thus the features selected in these sub-classifiers are normally also different. While some of the groups are rather "similar" to each other, it may lead to similar features selected, and thus some of the groups need to be deleted to meet the independent requirement according to the Dempster's rule.

For the groups considered to be "similar" to each other, only the group with the highest classification rate is kept; and all the other groups are deleted. For example, if 4 groups $G_{0n_1}, G_{0n_2}, G_{0n_3}, G_{0n_4}$ have classification rates as $R_{n_2} > R_{n_4} > R_{n_1} > R_{n_3}$, only the group G_{0n_2} is kept; and the other 3 groups are deleted.

Three criterions of judging "similar" groups are applied as follows:

a) The groups with the identical union of all the subsets.

For example, if the classification rates for the 4 groups

$$G_{0n_1} = \{E_1\} \cup \{E_2\} \cup \{E_3\}$$

$$G_{0n_2} = \{E_1\} \cup \{E_2, E_3\}$$

$$G_{0n_3} = \{E_2\} \cup \{E_1, E_3\}$$

$$G_{0n_4} = \{E_3\} \cup \{E_1, E_2\}$$

The union of all the subsets in these four groups is $\{E_1, E_2, E_3\}$, the four groups are considered as similar.

b) The groups with at least one common subset, and at least one common element in the other subsets.

For example, for the 2 groups

$$G_{0n_1} = \{E_1\} \cup \{E_2, E_3\}$$

$$G_{0n_2} = \{E_1\} \cup \{E_2, E_4\}$$

The subset $\{E_1\}$ is common for the 2 groups and the element E2 is common in the other subsets, the two groups are considered as similar.

c) The groups with all subsets included in the subsets of another group.

For example, for the 2 groups

$$G_{0n_1} = \{E_1, E_2\} \cup \{E_3, E_4\}$$

$$G_{0n_2} = \{E_1\} \cup \{E_4\}$$

The subset $\{E_1\}$ is included in the subset $\{E_1, E_2\}$, and the subset $\{E_4\}$ is included in the subset $\{E_3, E_4\}$, the two groups are considered as similar.

After the step of sieving of the groups, the kept groups are ordered in descending order according to the correct classification rate of the sub-classifiers as $G_1, G_2, \ldots, G_N$.

3) Step 3: Fusion between the sub-classifiers – application of the evidence theory

A step of fusion is applied to combine the sub-classifiers to get final decisions of the classification. The Dempster's rule of combination is applied to make data fusion between the groups of subsets. Each time of fusion is made between two groups of subset from groups G_1 and G_2. New fusion continues between the previous result subsets and the next group G until all the classes are separated in the subsets. The total number of fusions made K is defined as the fusion depth in this approach. An example of fusion of 2 groups is shown in Fig. 2, Fig. 3 showing the ambiguous classifier with a fusion of depth K.

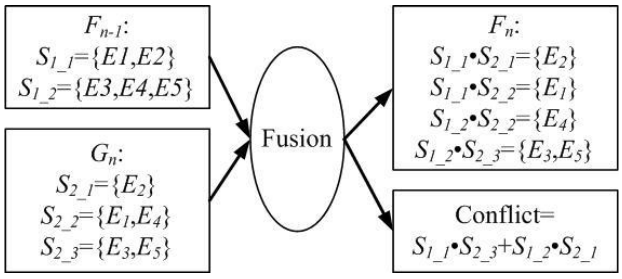


Fig. 2 Example of fusion of 2 groups of subsets

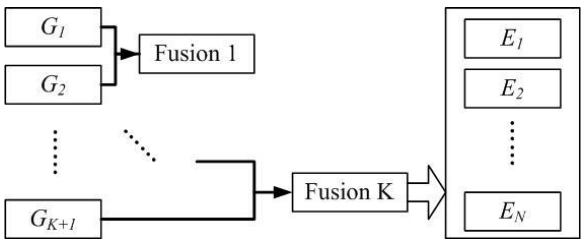


Fig. 3 Ambiguous classifier with fusion depth of K

The output of the classifier is the belief masses of the emotional states which satisfies

$$\sum_{i=1}^{N} m(E_i) = 1$$

To calculate the correct classification rate with the belief masses, two ways of evaluation are proposed in this approach. The first way, as the traditional classifiers with definite single judgment of recognized class for each sample, the class with the highest mass is judged as the recognized result. The second way considers multi possibilities of result. All the classes with masses larger than 0.3 are considered as possible recognition results.

3. Experiments and results

The ambiguous classifier proposed in the previous section has been evaluated on the problem of emotion classification in speech with Berlin dataset.

3.1 Dataset

The Berlin emotional speech database (Sendlmeier) is developed by Professor Sendlmeier and his fellows in Department of Communication Science, Institute for Speech and Communication, Berlin Technical University. This database contains speech samples from 5 actors and 5 actresses who pronounce 10 different sentences in German according to 7 kinds of emotions: anger, boredom, disgust, fear, happiness, sadness and neutral. There are totally 535 speech samples in this database, in which 302 speech samples are of female voice and 233 samples are of male voice. Because of the low number of disgust samples, this emotion has been omitted in our experiments. The length of the speech samples varies from 3 seconds to 8 seconds, and the sampling rate is 16 kHz.

Audio features used for the classification are those proposed in our previous work (Xiao, 2007). They include 5 groups of features to characterize the different audio properties of emotion: frequency features, energy features, harmonic features, Zipf features and MFCC (Mel Frequency Cepstral Coefficients) features.

4.2 Results

The automatically generated ambiguous classifier for the six classes in Berlin dataset is shown in Fig. 4. 5 sub-classifiers are selected to form 5 steps of fusions to get the belief masses of the 6 emotional states.

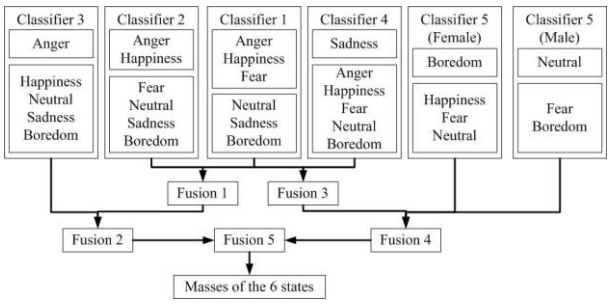


Fig. 4 Ambiguous classifier for Berlin dataset

The outputs of the ambiguous classifier are the belief masses of the emotions for each utterance. We proposed two ways of evaluations: traditional single judgment and

judgment of multiple possibilities. For the evaluation of single judgment, the utterances are classified as the emotional state with the highest belief mass among all the six emotions; and for the multiple judgment, all the emotional states with higher belief masses than a certain threshold are taken as possible results in the classification.

In our experiments, the threshold in the multiple judgments is set to 0.3. For the both evaluations, the classification results are obtained by the belief masses from same classifiers, only the way of applying the belief masses makes the difference as single or multiple judgments.

Fig. 5 shows the classification rates with single judgment, whereas Fig. 6 shows the classification rates with multiple judgments. In both of the figures, the curve "All samples (1)" refers to the case of classification for the utterances from mixed genders with a preliminary gender classification, and the curve "All samples (2)" refers to the case of classification for the utterances from mixed genders without gender classification. The error bars in the figures show the root mean square errors of the classification rates. From the difference between the curves "All samples (1)" and "All samples (2)", the gender classification shows obvious improvement in the overall classification performance for the mixed gender samples.

In the case of single judgment, the best result for female samples is 72.26%±1.82%; 74.62%±1.82% for male samples; 71.83%±1.80% for mixed genders with gender classification; and 61.04%±1.94% for mixed genders without gender classification.

In the case of multiple judgment, the best result for female samples is 75.81%±1.56%; 76.50%±1.69% for male samples; 72.32%±1.83% for mixed genders with gender classification; and 61.55%±1.82% for mixed genders without gender classification. The multiple labels in human testing on Berlin dataset are taken as ground truth.

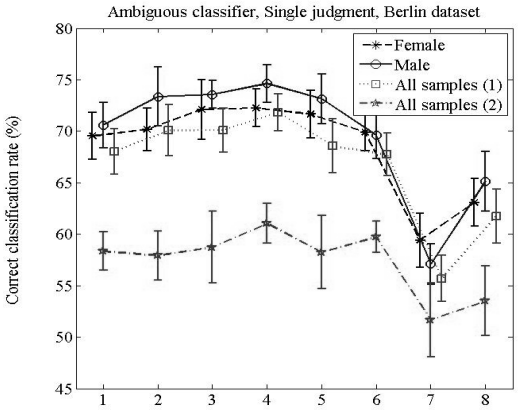


Fig. 5 Classification rate with single judgment

In the case of multiple judgment, the best result for female samples is 75.81%±1.56%; 76.50%±1.69% for male samples; 72.32%±1.83% for mixed genders with gender classification; and 61.55%±1.82% for mixed genders without gender classification. The multiple labels in human testing on Berlin dataset are taken as ground truth.

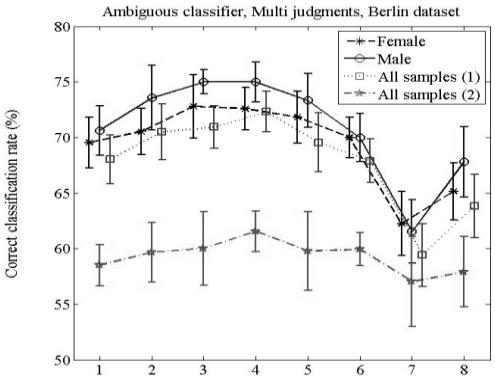


Fig. 6 Classification rate with multiple judgments

Distance between the confusion matrixes are obtained from human testing and automatic ambiguous classification with multiple judgments for the two genders respectively. The values of percentages are taken when calculating the distance. We consider the root mean square value of the difference between the matrices to present the distance:

$$D = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(CF_1(i,j) - CF_2(i,j))^2}{N^2}}$$

where N is the number of emotions, and CF_1 and CF_2 are the two confusion matrixes respectively. Meanwhile, the distances which evaluate each emotion are also calculated on each line of the confusion matrices as

$$D_i = \sqrt{\frac{\sum_{j=1}^{N}(CF_1(i,j) - CF_2(i,j))^2}{N}}$$

The distances between the results of human testing and automatic ambiguous classification with multiple judgments on Berlin dataset (presented on percentage) are listed in Table. 2.

	Whole matrix	E1	E2	E3	E4	E5	E6
Female	14.20	7.65	19.97	20.46	7.35	3.73	16.33
Male	11.50	12.59	13.88	11.63	9.74	9.86	10.72

Table. 2 Distance between the results of human testing and automatic ambiguous classification

From the distances listed in Table. 2, we can see that the judgments on emotional states neutral and sadness by the automatic approach is closer to human testing than the other emotions.

These results show that multiple emotion labels present common patterns in both human testing and machine recognizing. The utterances with original emotion labels as anger and happiness are frequently judged as anger or happiness at the same time, especially for the happy utterances in machine recognition in the ambiguous approach. Several emotions, such as happiness, fear and boredom, have relatively high chance to be recognized as neutral, especially for boredom. The speech samples in passive emotional states, sadness and boredom, are often judged as with both emotions in the ambiguous machine recognizing, while in human testing, only boredom is

frequently assigned with both emotions for male utterances, and sadness is seldom judged as also boredom.

Two rough summaries can be drawn from these results. First, these patterns in the ambiguous classifier with multiple judgments fit to the distribution of the emotional states in the dimension space: the emotions close to each other have more chance to be assigned simultaneously to a same utterance, and all the emotions might be assigned as with no particularly emotional states as neutral which locates in the center of the dimensional emotion space. Thus it proves that the mapping of the discrete emotions into dimensional space is reasonable in the emotion classification. Second, the similarity between the multi-label patterns in human testing and machine recognition shows the potential to simulate the human manner in subjective judgment of vocal emotions, even if only the absolute emotion labels are applied in the learning process of building the ambiguous classifiers.

4. Conclusion and Future Work

We proposed in this paper an automatic approach as ambiguous classification scheme of emotional speech which allows labeling the emotional utterances with multiple emotions is proposed and experimented on Berlin dataset.

As only a preliminary attempt on automatic ambiguous classification of emotions, our ambiguous classifier still needs to be greatly improved. The aim of proposing the ambiguous classifier of the emotional speech is to produce machine judgments to emotions as close as possible to the human ones. Two main aspects of improvement are needed in our future work.

First, in the learning process of the ambiguous classifier, the original emotional labels from the datasets with single emotion are still used as ground truth. Only the classification results are allowed to be labeled with multiple emotions, and the results on Berlin dataset are evaluated according to the multiple emotional labels obtained from human testing. In our future work, the ground truth of emotions for both the learning process and the evaluation will be modified to multiple emotional labels. Since the current public emotional speech datasets such as Berlin dataset are recorded for the analysis for typical emotions by professional actors, new dataset aiming the subjective vocal emotions will be needed.

Second, as we mentioned in the beginning of this paper, an emotion state is not only subjective which may lead to multiple possibilities in judgment, but also quite fuzzy and should be continuous which may lead to the different degree within the same emotional family, such as explosive happiness vs. calm happiness, or hot fury vs. cold anger. This continuity in the emotions is not yet considered in our work, and should be a topic of interest in the future investigates. Further to the two dimensional model with an arousal and an appraisal dimension used in our work, a more precise model of the position of the emotions in the dimensional space is needed for the research of continuous emotions, and other dimensions might be necessary to be introduced to make more reliable model of continuous emotion.

5. References

Scherer, K. R. (2002). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40, pp. 227--256.

Pereira, C. (2000). Dimensions of emotional meaning in speech, *Proceedings of the ISCA workshop on Speech and Emotion*, Newcastle, Northern Ireland, pp. 25--28.

Scherer, K. R. (2000). Psychological models of emotion. In: Borod, J. (Ed.), The neuropsychology of Emotion. Oxford University Press, Oxford/Newyork, pp. 137--162.

Thayer, R. E. (1989). The biopsychology of mood and arousal, *Oxford University Press*, 1989.

Sendlmeier et al. Berlin emotional speech database, available online at http://www.expressive-speech.net/.

Shafer, G. (1990). Axioms for probability and belief-function propagation. *Uncertainty in Artificial Intelligence*, 4, pp. 169--198.

Shafer, G. (1992). The Dempster-Shafer theory. *Encyclopedia of Artificial Intelligence, Second Edition*, Stuart C. Shapiro, editor. Wiley. pp. 330-331.

Fioretti, G. (2004). Evidence Theory: A Mathematical Framework for Unpredictable Hypotheses. *Metroeconomica*, (55:4), pp. 345--366.

Telmoudi, A.; Chakhar, S. (2004). Data fusion application from evidential databases as a support for decision making, *Information and Software Technology*, (46:8), pp. 547--555.

Xiao, Z.; Dellandrea E.; Dou, W. ; Chen, L. (2007). Automatic Hierarchical Classification of Emotional Speech. *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, Taichung, Taiwan, pp. 291--296.

Pudil, P.; Novovicova, J.; Kittler, J. (1994). Floating search methods in feature selection, *Pattern Recognition Letters* 15, pp 1119—112

Testimonials on emotions – a multimodal speech analysis

Gaëlle Ferré
LLING – Université de Nantes
Chemin de la Censive du Tertre, BP 81227
44312 NANTES Cedex 3
E-mail: gaelle.ferre@univ-nantes.fr

Abstract

The proposal in this pilot study is to present a prosodic and gestural analysis of testimonials on emotions in English and French, the premise being that while recollecting emotions, speakers give some representation of the emotions they convey. They actually shift from neutral to more emotional speech. The corpus described below is a series of podcasts in which ordinary people from different countries have been video recorded giving their opinion on issues that involve emotional speech. Three emotions have been considered: cold anger, sadness and happiness, and in terms of prosody, the paper concentrates on the intonation contours used by the speakers rather than on pitch span and F0 level. The annotation of gestures made with ELAN is based on a modified version of the MUMIN Coding Scheme (Allwood et al., 2005) and the conclusion of the paper is that speech is felt to be more emotional when a set of prosodic and gestural parameters is modified from more neutral speech, e.g. smiles or laughs do not in themselves convey happiness, but must be combined with other gestural and prosodic features to do so.

1. Introduction

This paper proposes a multimodal analysis of emotional speech compared to more neutral parts of discourse in podcast films recorded by the association "GoodPlanet" for a public exhibition directed by photographer Yann Arthus-Bertrand. The corpus description (see section 2) addresses the issue of the use of podcast recordings for research purposes. Among the films recorded by the association, I chose to analyze a collection of testimonies which where semantically linked with emotions, the underlying principle being that while speaking of their emotions, the speakers show a bit of these emotions in their expression. According to Ekman (1999:50) expressions "are part of an emotion; they are a sign that an emotion is occurring". This principle illustrates what Caffi & Janney, quoted in Plantin (2003:9), call emotional communication, i.e. "a type of spontaneous, unintentional leakage or bursting out of emotion in speech". And this is exactly what occurs in the videos, where speakers narrate some emotional experience to the public in a rather neutral tone, but at times, their emotions burst out in their intonation and facial expression. In section 4, I present a qualitative analysis of some salient prosodic and gestural features related to cold anger, sadness and happiness.

2. Corpus description

The corpus I worked on is a series of podcasts made available online by the association "GoodPlanet" headed by photographer Yann Arthus-Bertrand. The aim of the association's project "6 billion others" is to collect a series of testimonials from people around the world on a vast quantity of topics. Ordinary people from different parts of the world were asked to give their opinion on questions such as "what do you fear most?", "what has been the biggest joy in your life?, "on what occasion did you cry?", "what makes you particularly angry?" or "have you ever felt discrimination?". They were video-taped by a team of filmmakers of the association and the videos were edited into clips, the utterances said being inserted as a French subtitle into the clip. Each clip shows up to five or six people speaking different languages and expressing an opinion on a common topic, and lasts about 5 minutes. The project has been running for five years now and the final aim is to show the videos (more than 5000 interviews filmed in about 75 countries) in a free exhibition to be held in the Grand Palais in Paris in January and February 2009. The association has given consent for the video-clips to be used in a research project on emotions, provided the persons are treated with dignity and respect. This doesn't completely answer the issue on ethics, since such research work was not initially covered by the association's project, but the participants were recorded of their free will and accepted that the recordings be used in a public exhibition and displayed on the internet. No judgement will be made in this paper on the participants' physical appearance or on the relevance of their speech.

The major difficulty while working on podcasts is the issue of the naturalness of the data. The clips I worked on have been edited by professional filmmakers who obviously cut the sequences of hesitations and false starts always present in spontaneous data, but the resulting films are very close to natural data. It just means, in terms of prosody, that the clips can't be used for a study of speech rhythm, silent pauses and hesitation discourse markers, which is not the purpose of this paper. Another drawback of the clips is that the persons have been filmed during their testimonials but no unemotional speech of the same speakers has been recorded for comparison and although each clip contains testimonials on the same topic, speakers do not say exactly the same thing. Consequently, the object of the paper will be to study the shifts between neutral and emotional speech within each testimonial rather than comparing emotional speech between speakers.

When using podcasts, the major drawback lies in the often poor quality of the recordings. Most of the time, the recordings use a compression rate which is too low for a good image quality in terms of pixel number, a quality which is nevertheless necessary to allow even a manual

detection of fine facial details. This is not the case of the *GoodPlanet* podcasts which are of a high quality despite compression. The clips were edited in m4v, but I changed the codec into MPEG1 to be able to read the clips in ELAN (see section 8). Another drawback of podcasts is that speakers are usually filmed in action, with numerous zooms and close-ups which make it difficult for the analyst to annotate the movements of a particular part of the body, not always visible on the video. The testimonies of *GoodPlanet* are all filmed in the same condition, with an extreme close-up on the speaker's face. This is particularly convenient for the annotation of fine details of the face, although it means, of course, that other parts of the speakers' bodies are not visible on the screen. At last, a prosodic analysis of speech requires an extremely good quality of the sound of the recordings, and this is far from being the case in most podcasts. In the vast majority of podcasts, the sound-track is full of background noise, either music added to the initial recording, or surrounding noise during the recording which prevents any serious detection of prosodic parameters such as F0 or intensity. Another problem lies in the fact that in many recordings, several people speak in overlap and are recorded with a single microphone which also hinders the detection of prosodic parameters. *GoodPlanet* podcasts present none of these faults since the speakers are recorded professionally with no overlapping speech and background music or noise. The sound tracks of the clips are of a good quality and allow prosodic treatment of the data.

3. Annotation of the corpus

3.1 Phonology and prosody

The first step in the annotation process consisted in a prosodic annotation of about 15 mn of the corpus collection with PRAAT (see section 8). The sound files, extracted from the videos, were first converted into wave files to be readable in PRAAT and transcribed into current orthography in what could be termed intuitive prosodic groups. The translations made by the filmmakers could not be used since they didn't correspond to the exact words uttered by the speakers and were not aligned with the sound signal. From this initial transcription, words, syllables and phonemes were then annotated using the SAMPA convention of annotation, entirely manually for English and with the help of EasyAlign for French (see section 8). I then noted focal accents on a separate tier and on yet another tier, prosodic units (intermediate vs. intonational phrases, see Cruttenden, 1997:59-60) were noted as well as the general pitch contour on each phrase, using the contours described in Cruttenden (1997): high and low fall, high and low rise, rise-fall, fall-rise, and flat. In the cases where PRAAT displayed F0 detection errors, I used the Prosogram (see section 8) to obtain a better representation of the contour. The phonemic and prosodic annotation is shown in Figure 1 below:

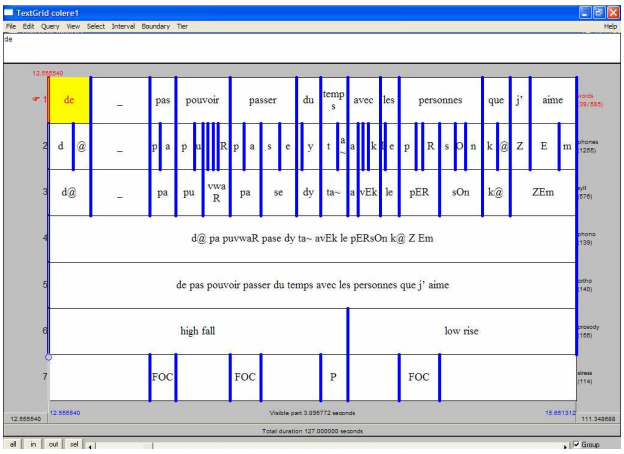


Figure 1: Textgrid of the orthographic, phonemic and prosodic annotation of the corpus in PRAAT.

3.2 Gestures

These annotations were then imported in ELAN, a tool for the annotation of video files (see section 8). For the annotation of gestures, I used an adapted version of the MUMIN coding scheme, fully described in Bertrand et al. (2007). The scheme was developed to describe gestures made in interactional data, which is quite poor in emotional load (Bouchard, 2000). Yet, the description of gestures using this scheme is extremely precise and a tier has been devoted to the annotation of emotions/attitude. I compared this scheme with the one recommended by the Network of Excellence HUMAINE and the annotation types of the two schemes are very close. Since the films were close-ups of the speakers' faces, I only annotated facial movements in a first step: gaze direction (front, left, right, up, down), head direction and head movements (shake, nod, tilt, beat...), eyebrow movements (frown, raise), mouth movements (smile, laugh...). In the file concerning sadness, I also added a track to annotate eye moisture (wet eyes) which was not initially thought of in the scheme, since this parameter seemed to play a role when the speaker displayed this emotion.

At last, in each file, I annotated the parts where speech seemed to be more loaded with emotion, simply stating the emotion (cold anger, sadness, happiness). This was made rather intuitively and by only one annotator for the moment. It is clear that in the future, several annotators will be needed for this annotation which is more subjective than the mere annotation of gestures, and their agreement checked by Kappa measurements.

Correlations of the different parameters were made with the help of the search function of ELAN.

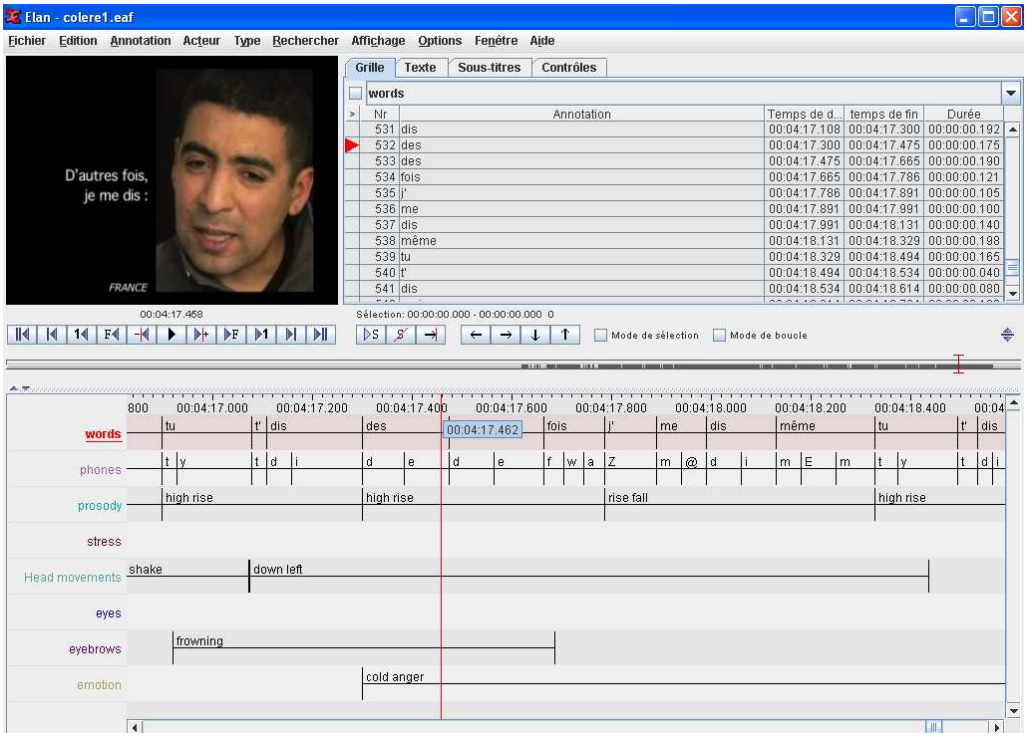


Figure 2: Snapshot of the annotation board in ELAN showing orthographic, phonemic, prosodic and gesture tracks.

4. Results and discussion

Although the files were too short to present any statistical analysis here, the results obtained are nevertheless interesting in a qualitative approach which may constitute the basis for further work.

4.1 Prosody

As stated earlier in this paper, the edited nature of the files does not allow work on rhythm (filled and silent pauses as well as speech rate) since parts of the file have probably been deleted and we do not have access to the original recordings. Yet, much can be done on pitch and stress. Previous studies have mostly described the links between emotional speech and speech span as well as F0 average height. Working on Japanese and French, Yamasaki (2004), for instance, found that positive emotions were rather perceived with rising F0, high average F0 being associated with gaiety. On the contrary, negative emotions were rather perceived when the F0 was falling, low average F0 being associated with sadness. She didn't however annotate contours from a phonological perspective. Devillers and Vasilescu (2004) studied the span and height of the intonation contour and found that in two negative emotions, fear and anger, the amplitude of the parameters depended upon subjects, but tended to be larger in emotional than in neutral speech. Both studies agree on the important weight of lexical/semantic content in the perception of emotional speech.

In this study, I concentrated on phonological F0 contours, comparing the distribution of contours for each speaker on the total testimony on one hand, and in correlation with emotions on the other hand. Results showed that there is a larger proportion of rising-falling contours in sadness as in the total testimony for this speaker. Happiness, on the contrary, was more correlated with low rising contours. This is in agreement with the findings of Yamasaki (op.cit.) associating lower intonation to negative emotions and higher intonation to positive ones. Looking at cold anger, however, an emotion which could be classified as negative, one finds that there is a higher proportion of high rises and rising-falling contours which do not correspond to Yamasaki's results. Yet, they confirm Devillers and Vasilescu's results (op. cit.) stating that the pitch span is larger in emotional speech, since I didn't find any high rising contours in declarative utterances in other contexts.

What makes sense however is that emotional speech implies a higher involvement of the speaker in his discourse and this is shown prosodically speaking with the use of a higher proportion of modulated tones such as the rising-falling contour. According to Morel & Danon-Boileau (1998), rising contours show a greater appeal to the co-participant in conversational data on the part of the speaker. In this context, which is not a conversation, the expectancy of empathy is nevertheless still present since the speakers may be said to convey a message to the public, represented by the eye of the camera. Emotional speech then can be described as conveying more involvement on the part of the speaker who also expects a greater involvement on the part of the addressee. Bagou (2001) and Plantin (2003) find that emotional involvement is regularly associated with emphasis in spontaneous speech. I also found that there was a high quantity of focal accents in the files I treated,

especially in the one showing cold anger. The focal accent is mainly realized through lengthening of word initial onset consonants. Yet, speakers in this file switched from rather neutral speech to speech showing cold anger and the proportion of focal accents was not higher in the speech parts perceived as more emotional. In other words, I could not find a direct link between prosodic emphasis and emotional speech, which shows that other parameters are needed for speech to be perceived as emotional. This is even reinforced by the gestural analysis (*Cf.* Analysis of "beats") in the next section.

4.2 Gestures

To follow my thread of thinking, the gesture annotation of the corpus showed that in the cold anger file, most focal accents were accompanied by a head beat of speakers (e.g. a rhythmic downward rapid movement of the head produced on accented syllables). These beats were not either associated with the perceived emotional parts of the file. This shows that a strong focal accent, marked both prosodically and gesturally, is not sufficient for speech to be perceived as loaded in emotional weight, although the utterance is not perceived as neutral, but as emphatic. In my opinion, emphasis and emotion are not necessarily linked.

Other gestures and physical properties were however typically associated with the three emotions under study. What appears immediately in Table 1 is that each gestural parameter allows a distinction between two emotions. For instance, head movements allow a distinction between sadness and happiness (shakes being associated with sadness and nods with happiness), whereas no particular head movement is found in correlation with cold anger.

	Gesture	Cold anger	Sadness	Happi-ness
Eyebrows	Raising		X	
	Frowning	X		
Head	Shakes		X	
	Nods			X
Gaze	Away		X	
	Front			X
Eyes	Narrowed eyes	X		
	Wet eyes		X	
Mouth	Laughs	X		X
	Smiles			X

Table 1: Relevant gestures and physical properties related to the three emotions.

Table 1 corroborates the results of Smith & Scott (1997), partially based on Darwin (1981), especially concerning eyebrow movements (raised eyebrows conveying sadness as opposed to frowns which convey cold anger).

I also found like Smith & Scott that laughs and smiles are associated with happiness.

Figure 3: Still pictures of speakers showing cold anger (left) and happiness (right).

What is different though is that laughs are also met in cold anger in this corpus, which is not however surprising. As stated in Bertrand et al. (2000), laughter reveals a distance of the speaker from his feelings. Laughter can then be met in other feelings than happiness, such as in cold anger and may be induced by the fact that the speaker suddenly realizes his greater involvement in a negative emotion from which he needs to take a distance. This is directly linked as well to the question of politeness evoked in Kerbrat-Orecchioni (2000:51) for whom « la civilité n'admet pas les manifestations émotionnelles intempestives et incontrôlées » and « la politesse et les rites sociaux ont précisément pour fonction principale de canaliser le flux affectif, de juguler les débordements émotionnels (...) ». This is probably also why speakers conveying sadness tend to look away from the camera when the emotion is too strong, whereas in happiness, they look straight at the camera, this latter emotion being a positive one. No regularity in gaze direction could be observed when the emotion conveyed is cold anger. At last, as far as eyes are concerned, this work corroborates Smith & Scott's findings (op. cit.) in which anger was associated with narrowed eyes. So far, the results for cold anger and happiness also corroborate the studies made by Cosnier & Huyghues-Despointes (2000) on the cognitive representation of emotions, a study in which they find that the mental representation of anger triggers movements in the subjects' eyes and eyebrows, whereas the representation of happiness triggers mouth movements. Smith & Scott did not test eye moisture, but in the file I worked on, it was obvious that sadness was perceived in a stronger way with greater eye moisture of the speaker.

Figure 4: Eye moisture in the perception of sadness.

This parameter is however not as reliable as the other ones though, since eye moisture on a video file may depend much on the eye colour of the speakers, as well as on the lighting used during the recording.

5. Conclusion

As a conclusion, I may say that this paper addressed several issues currently under discussion in the research community on emotions: firstly, concerning the data which lacks cruelly, it was shown that some podcast recordings may be used although work on such recordings cannot answer all the questions raised (but is this not the case of any corpus?). The corpus of the project "6 billion others" is of a quality which allows a prosodic and gestural treatment and I showed how speakers, through emotional semantic content, switch from rather neutral narration of their feelings to more emotionally involved speech.

The emotions conveyed were cold anger, sadness and happiness, which I understand as attenuated forms of anger, deep sorrow and joy. They were expressed through certain intonation contours: high rises and rising-falling contours for cold anger, rise falls for sadness and low rises for happiness. These preferred contours show more involvement on the part of the speaker who also seeks some empathy from the addressee. The speaker's involvement is however not linked to the presence of focal accents in this corpus.

In terms of gestures, emotions are conveyed with a set of different parameters: cold anger is associated with eyebrow frown, narrowed eyes and even laughter at times; sadness is associated with raised eyebrows, averted gaze, eye moisture and head shakes; happiness is associated with head nods, direct gaze, smiles and laughter. What both the prosodic and gestural analyses show is that prosody and gestures are not redundant in emotional speech, as already stated by Aubergé (2002), and that both participate in the perception of emotion. It also shows that analyses of video films are useful to complement studies based on pictures, since a laughing face in itself does not necessarily convey happiness, for instance. Yet, a still photograph of a laughing face would nevertheless be interpreted as a happy face without the context of the film and other parameters as already pointed out by Beavin Bavelas & Chovil (1997).

The limits of this study are also important. The quantity of data treated was quite small and an annotation of more recordings would be needed first to confirm the results of this pilot study, then to be really able to compare the results obtained on English and French which were treated here as equivalent.

6. Acknowledgements

I am particularly indebted to A.-L. Charriot and F. Gilard, producers of the project "6 billion others", for allowing me to use the recordings of the association "GoodPlanet", headed by photographer Yann Arthus-Bertrand. I hope the 2008 exhibition in Paris is a success and that useful work can be done in linguistics using their material. Any error in the description of the corpus or the aim of the association is mine.

7. References

Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C. & Paggio, P. (2005) The MUMIN Multimodal Coding Scheme, *NorFA yearbook 2005.* http://www.ling.gu.se/~jens/publications/B%20files/B70.pdf

Arthus-Bertrand, Y. (2003-2008). 6 billion others, a project of the Association "GoodPlanet", A.-L. Charriot - Directrice de Production & F. Gilard - Chargé de production, http://www.goodplanet.org/ (last access: 06-02-2008).

Aubergé, V. (2002). Prosodie et émotion. In *Actes des Deuxièmes Assises Nationales du GdR* I3, pp. 263—273,www.irit.fr/GDR-I3/fichiers/assises2002/papers/15-ProsodieEtEmotion.pdf (last access: 07-04-2008).

Bagou, O. (2001). Validation perceptive et réalisations acoustiques de l'implication emphatique dans la narration orale spontanée. *Cahiers de Linguistique Française*, 23, pp. 39--59.

Beavin Bavelas, J., Chovil, N. (1997). Faces in dialogue. In J. A. Russel & J.M. Fernandez-Dols (Eds.), *The Psychology of Facial Expression*. Cambridge, CUP, pp. 334--346.

Bertrand, R., Blache, P., Espesser, R., Ferré, G.; Meunier, C, Priego-Valverde, B., Rauzy, S. (2007). Le CID: Corpus of Interactional Data -protocoles, conventions, annotations-, *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence (TIPA)*, 25, pp. 25-55.

Bertrand, R., Matsangos, A., Périchon, B., Vion, R. (2000). L'observation et l'analyse des affects dans l'interaction. In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 169--182.

Bouchard, R. (2000). M'enfin !!! Des "petits mots" pour les "petites" émotions ? In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 223--238.

Chung, S.-J. (2001). Efficiency of the final part of the utterance in the communication of emotion. In C. Cavé, I. Guaïtella, S. Santi (Eds.), *Oralité et Gestualité (ORAGE) : "Interactions et comportements multimodaux dans la communication"*, Paris, L'Harmattan, pp. 183--189.

Cosnier, J., Huyghues-Despointes, S. (2000). Les mimiques du créateur, ou l'auto-référence des représentations affectives. In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 157--168.

Cruttenden, A. (1997). *Intonation*. Cambridge, CUP, second edition.

Darwin, C. (1981). *L'expression des émotions chez l'homme et les animaux.* Translated from English by S. Pozzi & R. Benoît. Bruxelles, Editions Complexe.

Devillers, L., Vasilescu, I. (2004) Détection des émotions à partir d'indices lexicaux, dialogiques et prosodiques dans le dialogue oral. In B. Bel & I. Marlien (Eds.), *XXVèmes Journées d'Étude sur la Parole*, AFCP, pp. 169--172.

Ekman, P. (1999). Emotional and conversational nonverbal signals. In L. S. Messing & R. Campbell (Eds.), *Gesture, Speech and Sign*. New York, Oxford University Press, pp. 45--55.
Kerbrat-Orecchioni, C. (2000). Quelle place pour les émotions dans la linguistique du XXe siècle ? Remarques et aperçus. In M. Doury, V. Traverso, C. Plantin (Eds.), *Les émotions dans les interactions*. Lyon, Presses Universitaires de Lyon, pp. 33--74.
Morel, M.-A., Danon-Boileau, L. (1998). *Grammaire de l'intonation : L'exemple du français*. Paris, Ophrys.
Network of Excellence HUMAINE. http://emotion-research.net/ (last access: 07-04-2008).
Plantin, C. (2003). Structures verbales de l'émotion parlée et de la parole émue. In J.-M. Colletta & A. Tcherkassof (Eds.), *Les émotions. Cognition, langage et développement*, Liège, Mardaga, pp. 97--130.
Smith, C.A., Scott H.S. (1997). A Componential Approach to the meaning of facial expressions. In J. A. Russel & J.M. Fernandez-Dols (Eds.), *The Psychology of Facial Expression*, Cambridge, CUP, pp. 229--254.
Yamasaki, H. (2004). Perception des émotions "positives" et "négatives" chez les auditeurs français et japonais à travers le contour de F0. In B. Bel & I. Marlien (Eds.), *XXVèmes Journées d'Étude sur la Parole*, AFCP, pp. 465--468.

8. Tools

Praat (Boersma P. and D. Weenick): A system for doing phonetics by computer. http://www.praat.org
EasyAlign (J.-P. Goldman): http://latlcui.unige.ch/phonetique/
Elan (H. Sloetjes): http://www.lat-mpi.eu/tools/elan/
Prosogram (P. Mertens): http://bach.arts.kuleuven.be/pmertens/prosogram/

Static vs. dynamic Gestural Icons of "*Feeling of Thinking"*

Vanpé Anne, Aubergé Véronique
GIPSA Lab - Institut de la Communication Parlée - UMR 5216 CNRS/INPG/UJF/Stendhal
Université Stendhal, Domaine Universitaire BP25, 38040 Grenoble Cedex 9, FRANCE
Tél. : +33 (0)4 76 82 41 9700 - Fax : +33 (0)4 76 82 43 35
E-mail: anne.vanpe@gipsa-lab.inpg.fr, veronique.auberge@gipsa-lab.inpg.fr

Abstract

Most studies concerning expressive communication concentrate on (visual/vocal/auditory) expressions of the speaker while he is talking. But information about what a speaker is doing while he is not talking is also important ("feedback", (Peters & al, 2005; Schröder & al, 2006)). We first tried to build an empirical methodology of ethograms for information about the (non)talker's mental or affective states, that we called 'Feeling of Thinking' (Loyau & Aubergé, 2006). Then we confronted some of the identified Gestural Icons with the perceptual validation of their relevance, in an association task with subject's self-annotation labels. We tested : (1) the static form of the Icons and their dynamic one; (2) three presentation conditions: whole face, upper part of the face only and lower part of the face only. The Icons were globally well identified, and can consequently be considered as relevant. Moreover, our results showed the importance of the dynamism for the 'Feeling of Thinking' perception and called additivity of the upper and lower parts of the face in terms of affective information into question.

1. Introduction

Most studies concerning expressive communication concentrate on (visual/vocal/auditory) expressions of the speaker while he is talking. But information about what a speaker is doing while he is not talking (out of turn-taking) is also meaningful ("feedback", (Peters & al, 2005; Schröder & al, 2006)). Information about these out of turn-taking activities may be expressed in the speaker's face, gestures, or interjectory vocalizations (Loyau, 2007; Loyau, & Aubergé, 2006), what we have called "Gestural Icon". Such expressions are important in human-machine interactions, as well as human-to-human ones.

We observed in our expressive corpus a large quantity of non-verbal information, even out of turn-taking. We want to investigate (1) to what extent this information conveys information about the mental or affective states of the (non)talker, (2) whether this information is static, as suggested by Ekman (1994), or is some information perceptible only dynamically or by the rhythmicity of the gestures.

In this proposal, we describe our (1) underlying theoretical framework, (2) our methodology, using an empirical ethogram labeling method to detect minimal Gestural Icons (Loyau, 2007; Loyau, & Aubergé, 2006), for studying expressive communication based on our work with emotional induction in human-machine interactions, (3) the perceptual validation method for ascertaining the relevance of isolated Gestural Icons that we identified, and (4) the results of these perceptual validation tests about the static form of the Gestural Icons and their dynamic ones.

2. "Feeling of Knowing" to "Feeling of Thinking"

The corpus is the audio and video expressive corpus "Sound Teacher of E-Wiz" (Aubergé, Audibert & Rilliard, 2006) from 17 subjects. The subjects are instructed to use a "revolutionary system" to learn the vowels of languages from around the world. They are seated in an insulated soundproof room in front of a computer screen. The format is a "Wizard of Oz" scenario, in which the subject thinks he is communicating with the computer, but in reality, it is a person (i.e., "wizard"). Subjects interact with the computer only by speaking; the computer communicates to the subject either by text and isolated vowels, or by execution of the requested task. Subjects are thus either reading, thinking or producing speech and they are not aware that they are being video-taped. It allows us to get authentic but controlled expressive expressions.

After each recording, subjects self-annotate their mental and emotional states during the experiment.

"Sound Teacher" provides minimal dialogue, in the sense that the subject knows that his turn-taking does not change the nature of the interaction, i.e., the computer does not interrupt the subject but waits for the subject to respond. Thus, there is no need for the subject to send feedback to his conversational partner (i.e., computer machine); nevertheless, during the "turn-taking of the machine", the subject expresses various mental and affective states.

In pilot studies, Swerts & Khramer (2005) described that in a task, in which the subject is asked to retrieve information from his memory, the subject often feels that he knows the answer, that it was stored in his memory, and that he would be able to retrieve it later but not at the moment. These studies show that while the subject is trying to recall information from his memory, he shows expressions revealing his mnesic process, called "Feeling of Knowing".

The Sound Teacher task reveals cognitive and affective processes broader than the mnesic task expressed in Feeling of Knowing. The expressions of these processes are gathered here in a generic phenomenology that we call

Feeling of Thinking (Loyau, 2007; Loyau, & Aubergé, 2006), the expressions of emotional and mental states.

3. Labeling of corpus: an ethogram methodology

Annotation of expressions is crucial in this study. We use three sources for annotation: (1) the 17 subjects each label their own mental and emotional states, (2) these are sporadically checked with perception experiments, and (3) two researchers who are not informed about the subjects' self-labelling, annotated the expressions by examining the audio and video recordings. The guidelines for the annotation aimed to arrive at minimal Gestural Icons as labels to the expressive corpus.

For this purpose, we apply a protocol stemming from ethology, by annotating our corpus using ethograms. Ethograms represent the inventory of species behavior. We annotated by using our competence as humans and by maintaining objectivity and avoiding semantic interpretations of expression. Every annotation is defined according to the criteria of direction, localization, speed, symmetry, repetitiveness, intensity and/or amplitude.

We use the research tool for annotating digital videos ANVIL 4.0[1], developed by Michael Kipp in DFKI to label our corpus. (cf. Figure 1).

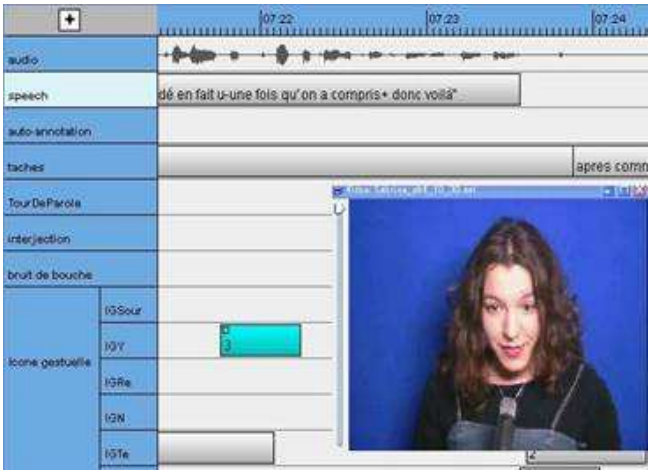


Figure 1: The corpus labeling using ANVIL

The corpus can thus be labelled objectively, not interpretively, by using primitive Gestural Icons for the chest movements, the face movements, and the vocal events.

4. Perceptual validity procedure

Some of our Gestural Icons were tested in order to perceptually validate them. Two identical perceptual tests were implemented: the first one with static forms of stimulus (pictures), the second one with the same Gestural Icons in their dynamic forms (videos) in order to investigate to what extent information of Feeling of Thinking is carried statically or dynamically.

The Icons were tested as to how well they matched the subjects' self-labels about their mental/emotional experiences (Vanpé & Aubergé, 2006). This was done by selecting two subjects with different personality profiles. One (Subject T) labelled herself as particularly stressed when told her answers were wrong; the other, (Subject S), laughed when she was told her correct answers were incorrect.

In order to select a characteristic and methodologically relevant subset of static and isolated Gestural Icons, we chose icons that had self-annotation labels representative of the two subjects' self-labeling of mental and affective states. Since the labels used were different for the two subjects, Gestural Icons of each one were evaluated in a distinct part of the test.

10 labels were therefore retained for subject T: "hesitant", "stressed", "ill at ease/worried", "anxious /oppressed", "at ease/more relaxed", "quiet/fine", "a bit lost/perplexed", "disappointed", "astonished", "concentrating"); and 9 for subject S ("not concentrating and feeling like laughing", "deriding my results", "listening with attention", ""holding over me" by the software", "stressed", "feeling like laughing and answering by chance", "concentrating and answering by chance", "concentrating" and "disappointed").

In order to assess Ekman's hypothesis about the importance of static information about different parts of the face (Ekman, 1994), three presentation conditions were tested: whole face ("whole"), upper part of the face only ("upper"), and lower part of the face only ("lower"), with a split according to a horizontal line located at the height of nose sidewalls. (cf.Figure 2)

Figure 2: Interface example – subject S stimulus in the lower condition

For every session of perceptual tests, each stimulus was presented once in each condition, for a total of 120 stimuli for subject T (10 labels * 4 sti./label * 3 conditions), and

[1] Complete documentation about ANVIL : Michael Kipp (2004), "Gesture Generation by Imitation - From Human Behavior to Computer Character Animation", Boca Raton, Florida: Dissertation.com; User manual at: http://www.dfki.de/~kipp/dissertation.html

54 for S (9 labels * 2 sti./label * 3 conditions). Stimuli were presented in a random order within "upper" and "lower" condition, while "whole" condition was always at the end, in order to avoid bias between subjects. Sixteen judges were presented the tests as closed choices among the self-annotation labels. Although observation time was not limited for static stimuli, dynamic stimuli (lasting 8 seconds) could be replayed.

5. Results

5.1 Dynamic Gestural Icons

5.1.1. Subject T

In the "whole" condition, the labels "concentrating", "anxious / oppressed", "astonished", "disappointed" and "at ease/more relaxed" were recognized above the chance level threshold by a factor of two. The other labels were not clearly recognized, but examination of the confusion matrix allows us to extract four meta-classes: "at ease/more relaxed" and "quiet/fine" (65.4%) ; "hesitant", "stressed", “ill at ease/worried", "a bit lost/perplexed", "disappointed" and "astonished" (51.7%); "anxious/ oppressed" (36.8%); "concentrating" (30.9%) (cf. Figure 3).

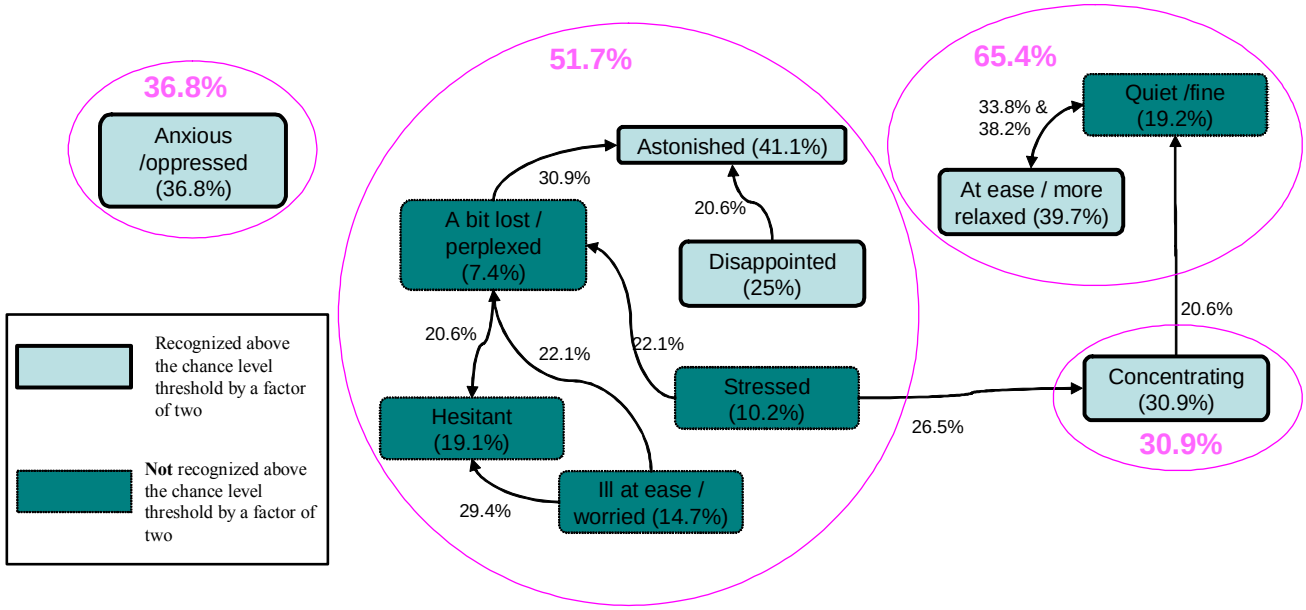


Figure 3: T. results with dynamic Gestural Icons in "whole" condition

5.1.2. Subject S

About subject S in the "whole" condition, the labels "not concentrating and feels like laughing", "feeling like laughing and answering by chance" and "concentrating" were recognized above the chance level threshold by a factor of two. Moreover, "concentrating" had good recognition in all conditions and attracted answers of other labels. We also extracted one meta-class: "feeling like laughing and answering by chance", "not concentrating and feeling like laughing" and "deriding her results" (identification rate from 51% in the "upper" condition to 76% in the "whole" condition). (cf. Figure 4)

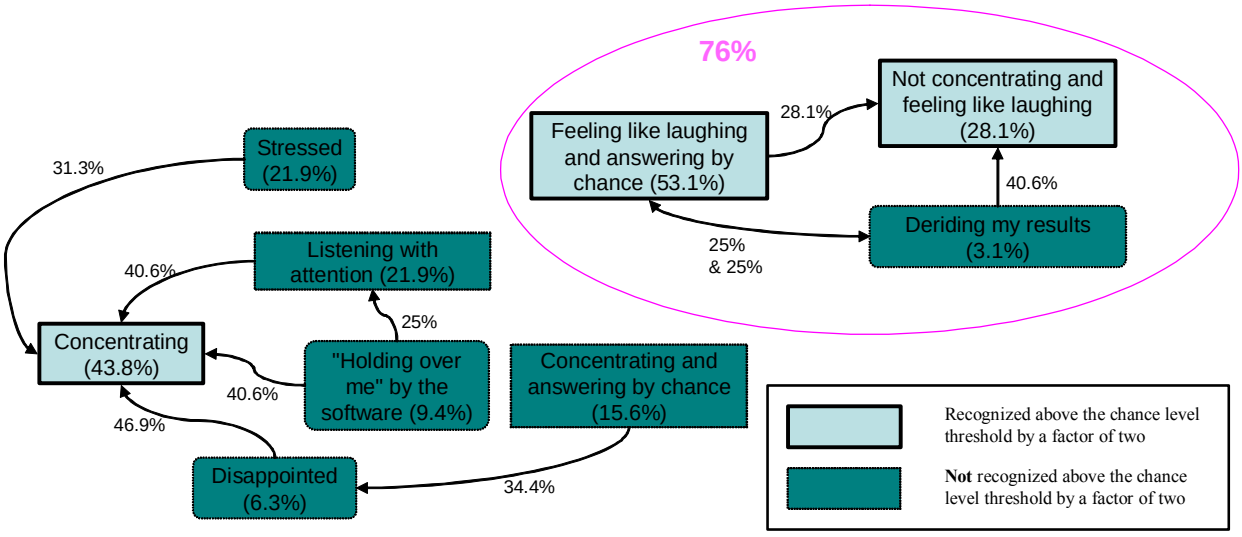


Figure 4: S. results with dynamic Gestural Icons in "whole" condition

5.2. Dynamic vs. Static comparison

Results for the dynamic Gestural Icons test were different from the static one (Vanpé & Aubergé, 2006).

5.2.1. Subject T

In the dynamic test results, we did not find exactly the same four meta-classes identified in the static test ("hesitant"/ "stressed"/ “ill at ease/worried"/ "anxious/oppressed"; "a bit lost/perplexed"/ "disappointed"/ "astonished"; "at ease/more relaxed"/ "quiet/fine"; and "concentrating"): "anxious/oppressed" was isolated and the first and the second one were merged. The labels "astonished" in all conditions, "anxious/oppressed" in the "whole" condition, and "hesitant" in the "upper" condition, were better recognized in the dynamic test, whereas "a bit lost/perplexed" in the "whole" condition, "anxious/ oppressed" in the "upper" condition and "disappointed" in the "lower" condition were better recognized in the static test.

5.2.2. Subject S

In the dynamic test results and in the "whole" condition, every label significantly differed from chance distribution (Khi-2, $p<0.01$, 8 ddl) (vs. in the static test results, it was the case of "not concentrating and feeling like laughing"). The labels "not concentrating and feeling like laughing" in the "whole" and "lower" conditions, and "concentrating" in the "upper" and "lower" conditions, were better recognized in the dynamic test, whereas "listening with attention" was better recognized in the static test in the "whole" condition.

Furthermore, the identified meta-class stayed the same in both test results.

6. Conclusion and prospective studies

The results of our study indicate that our Gestural Icons are perceptually relevant, both in their static form and in their dynamic ones. In addition, information given by the upper part of the face, and lower are not necessarily additive in terms of identification of labels, although this needs to be proven mathematically. Moreover it confirms that Ekman's FACS is not ecological, because expression of emotional states can't be reduced to a sum of Aus, minimal Gestural Icons, according to our terminology.

The upper/lower/whole condition and the dynamic or static nature of stimuli are two different parameters for identification of 'Feeling of Thinking' expressions. From static to dynamic results (Vanpé & Aubergé, 2006):

- Information about "concentrating" and "feeling like laughing and answering by chance" concentrates in the upper part of the face, as does information about "hesitant" and "listening with attention".
- The lower part of the face gives more information about "stressed" (information localized in the upper part of the face in the static test results), as does information about "quiet/fine", "at ease/more relaxed" and "not concentrating and feeling like laughing".
- Information about "disappointed" and "anxious / oppressed" (respectively localized in the lower and upper part of the face in the static test results) doesn’t seem to show a special localization, nor does information about "astonished".

In the future we wish to explore these results more to study the dynamics of the movements, since this seems to be important in some cases. For example, differences in dynamics may allow us to distinguish some labels. It also may be that the movement itself is important in some situations.

Moreover Carlier & Graff's works about upper-ranking tennismen (2006) suggest that gesture rythmicity (e.g., regularity and frequency), is a strong indicator of the subject's affective state. This needs to be checked using the Feeling of Thinking as an interaction task.

Regarding prospective studies, a first step would be to

focus on signal typology, ranging from static to dynamic. To measure co-ordination between the different modalities and to infer organization laws could then enlighten us about function and organization of rythmicity in emotional communication. We could also do perceptual evaluations with simulated icons using ECA. This would allow comparison of natural vs. synthetic movements in order to understand the essential cues of expressive speech facial gestures.

Based on these experiments, we would construct a model about the existing relation between multimodal expressions and the subject's mental and emotional states which could be then evaluated by using varied interaction-type data in order to validate or expand our model.

Finally the outcome of this work would be a simulation of this model with an augmented control of the virtual expressive agent GRETA, through a collaboration with C. Pelachaud and the LINC (Poggi & al, 2005). This simulation would be evaluated by usable tests, which would allow us to test our model as well as the ecological relevance of our simulations (cf. Figure 5).

Figure 5 : Subject T, E-wWiz corpus
vs. Greta (Pelachaud – LINC)

7. Acknowledgements

We warmly thank Donna Erickson for correcting our English. We also thank our expressive subjects and all judges who participated in the perceptual experiments.

8. References

Aubergé, V., Audibert, N., Rilliard, A. (2006). De E-Wiz à C-Clone. Recueil, modélisation et synthèse d'expressions authentiques. *Revue d'Intelligence Artificielle - "Interactions émotionnelles",* 20(4-5), pp.499—528.

Carlier, G., Graff, C. (2006). Unpredictability as a counter strategy: An analysis of elite matches. *Journal of Sport Sciences,* to be published.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115, 268-287.

Loyau, F. (2007). Expressions des états mentaux et émotionnels de l'humain en interaction : ébauches du "Feeling of Thinking". Thèse de Doctorat en Sciences Cognitives. Institut National Polytechnique de Grenoble.

Loyau, F., Aubergé, V. (2006). Expressions outside the talk turn: ethograms of the Feeling of Thinking. *5th LREC,* pp.47—50.

Peters, P., Pelachaud, C., Bevacqua, E., Mancini, M., Poggi, I. (2005). A model of attention and interest using gaze behavior. *IVA'05 International Working Conference on Intelligent Virtual Agents*, pp. 229-240.

Poggi, I., Pelachaud, C., de Rosis, F., Caroglio, V., de Carolis, B. (2005). GRETA. A Believable Embodied Conversational Agent. *Multimodal Intelligent Information Presentation*. O. Stock & M. Zancarano (Eds.), Kluwer, pp. 3--25.

Schröder, M., Heylen, D., Poggi, I. (2006). Perception of non-verbal emotional listener feedback. *Speech Prosody*, CD-Rom proceedings, SPS1-4-72.

Swerts, M., Khramer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language.* 53:1, pp. 81—94.

Vanpé, A., Aubergé, V. (2006). Pertinence perceptive d'Icônes Gestuelles du *Feeling of Thinking. Waca'06,* Toulouse, pp.55—59.

Cooperation Attitude in Negotiation Dialogs

Nicole Novielli
Department of Informatics
University of Bari
novielli@di.uniba.it

Peter Carnevale
Marshall School of Business
University of Southern California
peter.carnevale@marshall.usc.edu

Jonathan Gratch
Department of Computer Science
University of Southern California
gratch@ict.usc.edu

Abstract

We propose an annotation scheme for a corpus of negotiation dialogs that was collected in the scope of a study about the effect of negotiation attitudes and time pressure on dialog patterns.

1. Introduction

Affect has been shown to play an important role in negotiation dynamics: Kumar (1997) analyzed the role of positive and negative affect in bargaining; van Kleef et al. (2004) investigated the interpersonal effects of anger and happiness on computer-mediated negotiation. In particular, they considered the social consequences of emotions and their impact on the negotiators' strategic choices and their reaction to the opponent's affective state; Carnevale (2008) investigated the role of positive affect in simulated bilateral negotiation, with respect to the decision frame ("gain" vs. "loss").

Affective states widely vary in their duration, ranging from long-lasting features, such as personality traits, to short-term ones, such as emotions. Interpersonal stances are in the middle of this scale: at the beginning of the interaction they may be triggered by personality traits and they can stay unvaried over the whole duration of the interaction, unless significant events occur. This is especially true when the referred scenario is short-timed human-human interaction, such as negotiation dialogs in time pressure condition. Rather than considering individual emotions, we focused the research described in this paper on recognizing a particular aspect of interpersonal stance that influences the negotiators' behavior during interaction: cooperative vs competitive attitude. As we will see, this refers to the goals of the two parties involved in negotiation and how they behave to achieve them.

The final goal of this ongoing research is to investigate whether and how the cooperation attitude of the participants to negotiation dialogs, induced in an experimental study, can be recognized. The envisaged method for this recognition purpose is a combination of language analysis (at the individual move level) and dialog pattern classification techniques (Hidden Markov Models, HMM) (Charniak, 1993). In this short paper, we describe the corpus that will be used in this analysis and its annotation criteria and problems.

2. Conceptual Framework

A negotiation is a '*a discussion between two or more parties with the apparent aim of resolving a divergence of interests*' (Pruitt and Carnevale, 1993). It occurs whenever an economic transaction takes place or a dispute between goals is settled (Walton, 2005); a typical example is the labour negotiation scenario (Sycara 1989). Some recent studies (Carnevale and De Dreu, 2006; Carnevale and Pruitt, 1992) provide a deep insight on this phenomenon, from the *goals* and *motives* point of view. Schelling (1960) refers to negotiation as '*mixed motive*' interaction, meaning that the parties involved simultaneously experience the motivation to cooperate and compete with each other. In the basic types of dialog classification proposed by Walton (2005), negotiation dialogs are seen as originated by a conflict of interests of the parties involved. However, discrepancies on interests and goals do not necessarily produce a strictly competitive attitude in negotiators. Importance of goals with respect to the specific situation each party is in, together with their system of priorities or other factors as time pressure, might determine different kinds of attitudes and suggest negotiators to adopt different strategies (Carnevale and Lawler, 1986).

2.1 Cooperative vs competitive attitude

Competitive behavior occurs when parties assume a "win-lose" attitude, with strongly opposing interests. As a consequence, they could adopt a tough behavior (by highlighting unfairness of other's offers, putdowns) and coercive negotiation tactics aimed at forcing an advantage (e.g., threats), while excluding prenegotiation binding agreements (with exchange of priority information). To enhance the desired results, tactics which involve emotional ploys could be used, consistently with a strategic-choice perspective (van Kleef et al., 2004). On the contrary, in *cooperative* behavior a win-win situation is assumed, with the final goal to increase the joint gain. The resulting communication will be based on the hypothesis of

existence of common interests, benefits and needs, and will aim at building trust.

The features of competitive and cooperative attitudes we have described may be taken as cues for identifying this kind of behavior in negotiation dialogs. Typical *signs of cooperative attitude* are the accurate and honest exchange of information about own priorities and the spontaneous formulation of cooperative statements. To be successful, cooperative negotiators need to be skilled in clarifying similarities and differences in their individual goals and priorities and in trading, by proposing creative alternatives and selecting the best one, based on mutual acceptability. They will highlight consequences of a proposal for the other party which means showing understanding or interest for the other party's priorities ('*I know I can get more than that, but it cuts you down*') or evaluating the consequences for both ('*We are both maximizing our benefits*'). They will use 'positive' argumentation such as highlighting consequences of a proposal for the other party and will provide justification when making/rejecting a proposal. Skilled negotiators could also make use of humor as a technique of social influence (O'Quin and Aronoff, 1981), that is make a joke which does not involve a putdown of the other party ('*Think of my poor people! They aren't making any money here*').

On the contrary, in non-collaborative negotiation (Sycara, 1989; Carnevale and Lawler, 1986), it will be likely to see complaints about other's unfair offer ('*This is really lopsided*'), highlights of the other's contradiction ('*But you agreed with the other too*'), putdowns ('*Your workers are so stubborn!*'), self supporting statements ('*How about my proposal: 4c, 6/10ths and 40%? I think this is quite generous'*), threats ('*You don't want to be out of work…*') and warnings ('*If you don't agree with my proposal we will strike*').

2.2 The role of persuasion

Negotiation, persuasion and argumentation are close but not overlapping concepts. It is out of the scope of this paper to provide a clear definition of what are and what are not negotiation tactics, argument techniques and persuasion strategies: we are rather interested in defining a set of signs which can be used to detect the attitude displayed by the parties involved in negotiation processes. However, some preliminary clarification about the interrelationships among the three concepts is needed to justify our choice to introduce persuasion tags in our mark-up language. In analyzing agents' behavior in collaborative negotiation, Chu-Carroll and Carberry (1995) claim that '*argument is often taken to deal with conflicting opinions or beliefs, while negotiation deals with conflicting goals or interests*'. If negotiation is seen as a process aimed at defining an agreement on the two parties' conflicting goals (Walton, 2005), persuasion phases are easily embedded in these processes (Walton, 2005). The inverse is also possible: Wells and Reed (2006) show how people decide to embed negotiation sub-dialogs in persuasion ones, when they realize they are unable to change the goals of their opponent. Distinguishing between bargaining and argumentation can be difficult (Chu-Carroll and Carberry, 1995) and some authors claim about the existence of mixed types of dialogs (Walton, 2005).

3. The corpus

The corpus we used for this study was collected in the scope of a study about the effects of time pressure (Carnevale and Lawler, 1986). The experimental setting was designed to be a 2x2x2 study where the variables involved were time pressure (high or low), attitude of the negotiators (cooperative vs competitive) and their gender. The subjects involved in the experiment where asked to play the roles of union and management representatives, in a labor negotiation scenario. They were asked to negotiate on wages, medical plan and vacation and were given an issue schedule where their priorities were expressed in points assigned to each configuration of the three parameters. The subjects were told that the final value of the agreement reached would have been converted into real money. Subjects were privately provided of instructions about time pressure and orientation. The time pressure condition was simulated by giving a temporal deadline of five minutes while, to manipulate their orientation, subjects were explicitly instructed to behave as to reach an integrative agreement (cooperative attitude) or to maximize their own gain (competitive attitude). Therefore, people involved in the study modified their attitude according to spontaneous adaptation to the environment condition (high vs low time pressure) but also because they were instructed by the experimenter on how to simulate either a cooperative or competitive attitude. From this point of view, the corpus is half way in between spontaneous emotion corpora and acted ones.

	Number of dialogs available		Average number of moves	
	High time pressure	Low time pressure	High time pressure	Low time pressure
Cooperative attitude	12	12	32	66
Competitive attitude	11	11	29	86

Table 1: The time pressure corpus.

Same-gender subjects (24 male vs 24 female) were distributed equally and randomly through the four

combination of the two modalities. Table1 summarizes the distribution of dialogs in the four modalities.
Dialogs analysed in this study are the transcripts of audio recording of the negotiation experiments. We annotated, overall, 2433 moves. By 'move' we refer to the single turn performed by each party in the scope of a dialog exchange. Our final goal, in fact, is to model the attitude of both parties involved in the negotiation. For this reason, we think the annotation should be done at the single move level rather than at the level of coupled pairs composing the dialog exchange (tab. 2)

Speaker	Transcript	Possible unit of annotation	
1	*You want wages at what?*	Speaker move	Dialog exchange (complete turn)
2	*At 7, the original*	Speaker move	

Table 2: Possible annotation units.

4. Markup language

The first question to be answered when deciding to annotate a new corpus is whether to define an ad hoc scheme or to use an existing one. Several schemes have been proposed for coding bargaining processes (Goering, 1997): the main advantage of using an existing coding system is the possibility of comparative analysis with previous studies in the same domain. However, Weingart et al (2004) argue in favour of defining ad hoc annotation schemes according to goals researchers want to achieve and to the intrinsic features of the corpus to be annotated.
The first issue to be addressed when approaching the definition of a coding scheme is what are the relevant features of the phenomenon to be annotated (Craggs, 2003). Theoretical background and domain knowledge help in formulating a first sketch of the annotation language; inspection of the corpus (with computation of the frequency of labels in the dataset) should be addressed in further iterative revision steps, towards the definition of the final language. In negotiation dialogs, in particular, it must be decided what types of behaviour are theoretically relevant in the study and what are the cues through which this behaviour is shown. Also, the collection modality affects the kind of signs that may be looked for: spoken corpora provide information about prosody and other acoustic features; audio-visual data make possible the usage of body measures; transcribed or written corpora (as in our case) only allow linguistic analysis. Another critical issue to be addressed is the definition of the unit of annotation. It is very important to have a clear idea, since the beginning, of what the long-term goal of the research will be, so as to avoid loss of relevant information (too large units of annotation). This is particularly true when attempting to annotate a subjective phenomenon such as affective states. The general approach is to allow redundancy (e.g. by annotating single word units as in Batliner et al. 2003) and overlapping among tags: aggregation is always possible, a posteriori, while with further specification of tags researchers would introduce a subjective bias in the annotation results.

4.1 Definition of codes

The annotation language we defined extends the coding scheme used in a related study (Carnevale and Lawler, 1986). The core of this language includes *domain* tags: making/accepting/rejecting a proposal, bargaining and soliciting a reaction). Some of the existing tags were grouped into the category of those denoting, in particular, *cooperative attitude*: cooperative statement and exchange of priority information; finally, the language is extended with tags denoting *persuasion attempts.*
When dealing with the annotation of our corpus, we had to consider that consequences of time pressure condition in negotiator's attitude were quite natural since they had to spontaneously adapt to the high/low time pressure condition, which was simulated by adopting a time deadline. On the contrary, subjects were explicitly instructed to simulate a cooperative or competitive attitude by using written guidelines provided just before the experiment started. In this sense, our corpus is half way between spontaneous and acted data and this should be considered when summarizing results of the annotation experiment and approaching the attitude model learning. To this aim, *faithfulness to role* tags were added to evaluate how much the subject involved in the experiment behaved in a way which is close to 'real' negotiation interaction. This tag is used whenever a subject makes a comment drawn from the context ('*My constituents have been hard workers and deserve a higher salary*') or related to the experimental setting ('*We can get more points by doing this way*'). In the first case, the tag value indicates that the subject is behaving according to the role assigned, while the second one shows a situation in which the subject does not seem to be really involved in the negotiation task, as he explicitly refers to the experimental settings or to the instruction received, while interacting with the other party. This tag can be used to assess the validity of conclusions drawn from the analysis of this corpus and to assess the quality of data collected by asking people to 'act' as they were adopting a particular attitude (cooperative vs. conflicting). The complete set of codes, with definitions and examples, is provided in table 3. Please note that the level of generality of the four groups of tags is not necessarily the same, in order to allow, in the next future, several approaches of analysis at both linguistic (single dialogue turns as units of analysis) and pragmatic level (overall dialogue pattern). In particular, domain tags are

useful because they enable us to describe the actual evolution of the negotiation proposal, regardless of the behaviour subjects involved in the negotiation experiment are showing. It is reasonable to assume, in fact, discrepancies between the shown attitude (that we can recognize by looking at linguistic features of dialog turns) and the actual one (that we might analyze by looking at the evolution of the negotiation dynamics).

Group	Signs with definition	Value	Examples
Cooperative attitude	**Cooperative statement:** speaking positively about a mutually acceptable solution or about allowing both sides to do well.	**Yes**	*'I guess maybe we should start with where we can agree. According to this we are trying to maximize both our own and our partner's point ratings'* (Cooperative statement = 'Yes')
		No	
	Exchanging priority information: any honest exchange/request of exchange between negotiators about their priorities, according to the information provided by experimented and present on their issue sheets.	**Request info**	*'Let's exchange information about our point values'*
		Give info	*'The most important of the three issues for my point of view is the medical plan'*
Domain tags	**Making a proposal:** making an offer, either by simply presenting the proposal or also by supporting it with argumentation	**Simple proposal:** (single or multi-issue focus)	*-'Let's make a 5% on wages' (single focus)* *-'5c increase in wages 4/10 and 60%' (multi-issue)*
		By using **persuasive information** and argumentation	when using this code, raters had also to specify a value for the 'Persuasion Attempts' code (refer to that code for examples)
	Accepting a proposal: by either demonstrating interest in an offer without accepting it (**Open option**) or explicitly accepting it (**Offer acceptance**)	**Open option**	*'An option that we can keep open'*
		Offer acceptance	*'Okay'* (after a proposal from the other party)
	Rejecting a proposal: either by expressing disagreement in a **Polite way** or by expressing total unwillingness to make further concessions (**with heavy or impolite commitment**)	**Polite way**	*'I don't like that idea'*
		With heavy or impolite commitment	*'That's totally out of question'*
		By using **persuasive information** and argumentation	when using this code, raters had also to specify a value for the 'Persuasion Attempts' code (refer to that code for examples)
	Bargaining: speaker makes a proposal which suggests giving up on one issue in return for gaining on another issue	**Yes**	*'If we go down on vacation, will you go up on something else?'* (Bargaining = 'Yes')
		No	
Persuasion tags	**Highlighting consequences of a proposal for the other party:** statements which indicate understanding/interest in knowing the other's party priorities, joint evaluation of consequences,	**For both**	'*In this way we are both maximizing our benefits'*
		For the other	*'I know I can get more than that but it cuts you down'*
	Persuasion attempts	**Signs of cooperative attitude/real persuasion attempts**	Humor
		Signs of competitive attitude/making tactical use of power	Highlighting the other contradiction
			Complaining about other's unfair offer
			Self supporting statements
			Threats and warnings
Faithfulness to role	**Comment drawn from context**: any argument referring to surrounding social or economic structure which is used by one party to gain a concession from another		*My constituents have been hard workers*
	Argument related to the experimental setting: the speaker clearly refers to his point balance, instead of using arguments drawn from the context		*We can get more points this way*
Other	**Soliciting a reaction:** speaker requests the other's reaction (feelings or thoughts concerning an offer or general suggestion)	**Yes / No**	*'Let's hurry up and finish'* (Soliciting a reaction ='Yes')
	Asking a question		*'Do you want to start with wages'*
	Answering		*'Yes, let's start with wages'*

Table 3: The mark-up language.

It is possible, in fact, that parties were just pretending to be cooperative, for example by exchanging priority information. Actually, this could be a tactic adopted by skilled competitive negotiators: while collaborative agents really take into account others' beliefs, also in order to decide whether to revise their own ones and reach an agreement (Chu-Carroll and Carberry, 1995), competitive agents could do so to discover weak points in the other's system of beliefs and goals and to attack them with arguments or emotional tactics. Including these tags in the coding scheme leave us the door open to future investigation in this direction (e.g. aggressive verbal behavior as an emotional tactic, when a cooperative strategy is actually being adopted).

4.2 Labelling units

Decision about which unit of annotation should be used must be linked to the research question to be answered and to the further analysis that researchers intend to conduct on the annotated dataset. As we said, our long term goal is to learn an Hidden Markov Model which enables us to recognize negotiators attitude (such as, in this case, cooperative vs competitive behaviour) as in Martalò et al. (2008). For this reason, we found it relevant to label the dialogs at the entire dialog turn level. Some authors claim the possibility of letting raters free to manually divide the corpus into annotation units. We believe that this would make the annotation of subjective phenomena less reliable. The state of the art on this subject (Weingart, 2004 Craggs, 2003; de Rosis et al., 2006; de Rosis et al., 2007) suggest us to use objectively defined units of annotation (as dialog turns are). This allows us to evaluate the inter-rater level of agreement by using an index such as the observed agreement or Cohen's Kappa, which is recognized as a valid measure of interpretation reliability in the computational linguistics community.

4.3 The labelling experiment

A labelling experiment was conducted at USC: the three raters were all English native speakers and were provided of the complete corpus of transcribed data and of an annotation manual which explained in detail the meaning of each tag, by also providing examples. After an individual short training of about ten minutes, where raters were free to ask questions about how to conduct the labelling, they were asked to rate dialog moves independently. Multiple annotations were allowed because of partial overlapping in the semantic of some of the tags.

After summarization of results, every move received one or more codes according to a majority voting criterion (at least two over three raters agreeing on the value of a code). The main problem related to this approach is the probability of having no tags for some turn, when majority agreement is not reached: on the contrary, since our final aim is to train an HMM model for predicting the overall attitude of the negotiators during interaction, we need to give a code to all turns in the training set. Sparse data are also a relevant problem in model learning. For these reasons, we revised the corpus annotation by compacting the initial tags into fewer classes, according to the final recognition goal and to the semantic of codes (see table 4). This table shows the distribution of labels in the annotated corpus and provides values for the observed agreement and Kappa among raters (we didn't report the signs for which frequencies was zero). Which index best fits the description of the inter-rater agreement is still an open discussion in the computational linguistic community (Craggs and McGee Wood, 2004): while the observed agreement doesn't suffer from the unequal distribution of labels, Kappa provides a chance corrected measure of the agreement. Our results seem to confirm this issue: the signs for which we have the highest differences between the first measure (a percentage agreement index) and Kappa, in fact, are those with the lower frequency in the corpus (e.g. *Exchanging priority information*).

Sign	Frequency	Observed Agreement	Kappa
Cooperative statement	17 %	.79	.24
Exch. priority information	4.4 %	.83	.11
Making proposal	18 %	.81	.41
Accept proposal	5.1 %	.91	.35
Reject proposal	4.1 %	.93	.46
Bargaining	8.8 %	.83	.26
Soliciting reaction	7.9 %	.85	.27
Persuasion attempt	7.6 %	.83	.21
Faithfulness to role	5.8 %	.86	.26

Table 4: results of the annotation experiment.

5. Conclusion and future work

This contribution is a preliminary statement of the direction in which we are moving in our study about recognizing cooperation attitude in negotiation dialogs. By tagging our corpus, we made a first step towards preparing the dataset to be used to train our recognition model. Once again we learnt, from this experience, that the markup language is a compromise between the dimension of the corpus available, the data analysis goals, the methods that will be used in this analysis and the complexity of the problem under study, the actual features of the corpus used and how they can be described with the aim of building a model of those affective states which are relevant for the domain of application.

The limited dimension of our corpus and the unequal distribution of codes caused a low Kappa, even after aggregation of some of the tags according to their frequencies and to their semantic. This suggested us to carefully revise the results of the annotation experiment before starting any model learning phase. In particular, we must take into account the hybrid nature of the affect expressed in our corpus: on one hand, time pressure

condition and the perspective of higher money gain in case of successful integrative bargaining could promote a spontaneous adaptation of the negotiator's attitude; on the other end, the only way for inducing people to differentiate their attitude (cooperative vs. competitive) was the usage of written guidelines. Since no assessment was conducted on personality traits, we cannot be completely sure of how the same subject would behave in a real-life negotiation scenario, and how their permanent features would affect the interpersonal stance occurring in real-life situations. This suggests us to be careful in the preparation of the dataset for further analysis, also according to the information provided by the 'faithfulness to the role' code.

6. Acknowledgements

This work was jointly financed by HUMAINE, the European Human-Machine Interaction Network on Emotions (EC Contract 507422) and the Internship programme of ICT-USC. We warmly acknowledge the raters who took part in the annotation experiment at ICT and Fiorella de Rosis for her precious suggestions.

7. References

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E. (2003). How to Find Trouble in Communication, In: *Speech Communication 40*, pp. 117--143

Carnevale, P.J. (2008). Positive affect and decision frame in negotiation. In *Group Decision and Negotiation*, 17. Springer, pp. 51--63.

Carnevale, P.J., De Dreu, C.K.W. (2006). Motive: The negotiator's raison d'etre. In L. L. Thompson & J. M. Brett (Eds.), *Negotiation theory and research*. New York: Psychology Press, pp. 55--76.

Carnevale, P.J., Lawler, E.J. (1986). Time pressure and the development of integrative agreements in bilateral negotiation. In *Journal of Conflict Resolution*, 30. pp. 636--659.

Carnevale, P.J., Pruitt, D.G. (1992). Negotiation and mediation. *Annual review of psychology*, 43, 531-582

Chu-Carroll, J., Carberry, S. (1995). Response generation in collaborative negotiation. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 136--143.

Charniak, E. (1993). *Statistical language learning*. The MIT Press.

Craggs, R., Wood, M.M., (2003). Annotating emotion in dialogue. In *Proceedings of the Fourth SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, pp. 218-225.

Craggs, R., Wood, M.M. (2004). A two dimensional annotation scheme for emotion in dialogue. *In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. Stanford University.

de Rosis, F., Novielli, N., Carofiglio, V., De Carolis, B. (2006). User modeling and adaptation in health promotion dialogs with an animated character. In *Journal of Biomedical Informatics*, 39(5), pp. 514--53.

de Rosis, F., Batliner, A., Novielli, N., Steidl, S. (2007). 'You are soo cool Valentina!' Recognizing social attitude in speech-based dialogues with an ECA. In: *Proceedings of ACII 2007*, Lisbon. Springer, pp. 179--190.

Goering, E. M. (1997). Integration Versus Distribution in Contract Negotiations: An Interaction Analysis of Strategy Use. In *Journal of Business Communication*, 34, (4), pp. 383--400.

Kumar, R. (1997). The Role of Affect in Negotiations An Integrative Overview. In *The Journal of Applied Behavioral Science*, 33 (1), pp. 84--100.

Martalò, A., Novielli, N., de Rosis, F. (2008) Attitude display indialogue patterns. In *Proceedings of AISB '08, Symposium on Affective Language in Human and Machine*.

O'Quin, K., Aronoff, J. (1981).Humor as a Technique of Social Influence. In *Social Psychology Quarterly*, 44, (4). American Sociological Association, pp. 349--357.

Pruitt, D.G., Carnevale, P.J. (1993). *Negotiation in social conflict. Buckingham.* England: Open University Press.

Schelling, (1960). *The strategy of conflict*. Boston, MA: Harvard University Press.

Sycara, k. (1989) Argumentation: Planning other agents' plans. In *Proceedings of the 11th international joint conference on artificial intelligence*, pp. 517--523.

Van Kleef, G. A., De Dreu, C.K.W., Manstead, A.S.R. (2004). The Interpersonal Effects of Anger and Happiness in Negotiations. In *Journal of Personality and Social Psychology*, 86 (1), pp. 57--76.

Walton, D. C. (2005). How To Evaluate Argumentation Using Schemes, Diagrams, Critical Questions And Dialogues. In *Scoms: Argumentation in Dialogic Interactions*, pp. 51--74.

Weingart, L.R, Olekalns, M., Smith, P.L. (2004). Quantitative Coding of Negotiation Behavior. In *International Negotiation*, 9 (3). Martinus Nijhoff Publishers, pp. 441--456.

Wells, S. and Reed, C.A. (2006). Knowing when to bargain. In P. E. Dunne and T. J. M. Bench-Capon (Eds), *Proceedings of COMMA 2006, Working Notes of the 6th Workshop on Computational Models of Natural Argument*, pp.235--246.

Developing Affective Intelligence For An Interactive Installation: Insights From A Design Process

Lassi A. Liikkanen, Eero Huvio, Rodolfo Samperio
Helsinki Institute for Information Technology (HIIT)
Helsinki University of Technology and
University of Helsinki
P.O. Box 9800, FI-02015 TKK, Finland
E-mail: {firstname.surname}@hiit.fi

Eero Väyrynen, Tapio Seppänen
University of Oulu
Information Processing Laboratory
P.O. Box 4500,
FI-90014 University of Oulu, Finland
E-mail: {firstname.surname}@ee.oulu.fi

ABSTRACT

This paper documents a case study from the development of an affective application called PuppetWall, which is an interactive installation built upon the puppeteering metaphor. It is designed to react to user expressions and visualize them on a large multitouch screen. We present an outline of the system and a review of comparable applications. We describe our initial design efforts in implementing emotion recognition using speech and a novel way of using affective information to control the application. Based an initial user test, we show how users try to exploit the system by eliciting various vocal expressions. We conclude our presentation by examining the lessons learned from this design iteration, focusing on the auditory cues available and the implementation of interactive features.

1. INTRODUCTION

Natural interfaces are a new trend in human-computer interaction. They enable users to interact with advanced multimodal applications in a more embodied way, for instance, using hand gestures, bodily movements, or speech to interface them. Another line of progress concerns the tracking of emotional and expressive cues. These interface technologies provide tools to build innovative applications, where media is not just created and browsed in the traditional way, but is also enlivened in real time using multimodal inputs and emotional intelligence. One scenario is to create new formats which encourage users to animate media and co-create narratives. The development of such applications requires answering several open research questions. These include the identification of suitable input modalities, investigating the expressive features in each modality, and creating interaction loops that motivate users in self-expression.

The technology required to detect emotions and implement affective interaction has just recently started to bloom (Cowie et al., 2001). The slow progress maybe due to the fact that humans naturally use multimodal information, including semantics, to discover emotions and these aspects have long been a stumbling stone for artificial intelligence. This far computational methods have often been relying on a single modality in a restricted, context insensitive way and multimodal fusion in emotion decoding seems to be ahead of us (but see Paleari & Lisetti, 2006; Pantic & Rothkrantz, 2003). Despite the challenge, we believe that the time is right to begin explorations into affective interaction, even if the recognition technology is still being refined.

One of the oldest tracks in emotion research concerns vocal expressions (Scherer, 2003). We also adapt emotional speech as a starting point for developing emotional intelligence for PuppetWall, an interactive, affective installation. We try to address the question of what kind of auditory cues would best serve the interests of system development. The role of corpus is known to be crucial, but excluding some generic guidelines (Ververidis & Kotropoulos, 2006), there are no firm rules for corpus acquisition, and much has be to be done by (expensive) trial and error. We will briefly present of the technical details of our application, introduce some related applications, and provide results and conclusions from a small user study.

2. RELATED WORK

Puppeteering metaphor has been previously explored in digital domain to some extent. Interfaces for actors to control virtual characters have been developed, for instance a data glove and a custom sign language to control the behavior of a digital puppet (Camurri et al., 2005). Chinese shadow puppetry has been implemented in a more extensive system called I-Shadows (Paiva et al., 2006). That installation allowed children to create stories for an audience. These applications did not acknowledged users' emotions, but emotionally intelligent systems and their drafts do exist in other domains. McQuiggan and Lester (2007) have designed agents that are able to empathically respond to six emotions that match the gaming situation of a user. AffectivePainting (Shugrina et al., 2006) supports self-expression by adapting in real time to the perceived emotional state of a viewer which is recognized from his or her facial expressions.

Cavazza et al. (2004) introduce a prototype of multimodal acting in mixed reality interactive storytelling, in which the position, the attitude, and the gestures of spectators are monitored and influence the development of the story. Camurri et al. (2005) propose multisensory integrated expressive environments as a framework for performing arts

and culture oriented mixed-reality applications. They report an example in which actress' lips and face movements being tracked by the EyesWeb system and her voice is processed in real-time to control music. To sum up, many interesting approaches exist, but none of them include both full-blown interactivity and affective features.

3. PUPPETWALL

PuppetWall is a multi-user installation for collective, emotionally augmented interaction. It is based on the concept of a traditional puppet theater. Users control puppets and other elements of the application on a large multitouch screen with hand movements. In this section we provide condensed specifications of the current version of PuppetWall (for details see Liikkanen et al. 2008).

3.1 System Overview

PuppetWall system consists of several inputs for explicit and implicit interaction. MagicWands are tracked in 3D to capture hand gestures. They provide an intuitive and simple way to animate characters. Voice input is used to feed a speech classifier to analyze users' emotions, and a large multitouch screen is used for direct interaction and projecting the visuals of the application. The system architecture is visualized in Figure 1.

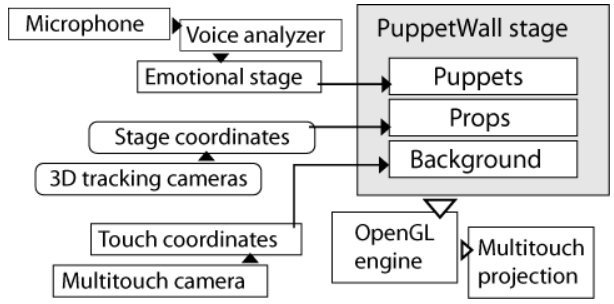


Figure 1: The architecture of PuppetWall application

This prototype runs on a single Linux workstation equipped with a 3D accelerated graphics card, which is utilized by a custom made 3D engine. The PC has two microphone inputs and three additional FireWire ports. The visuals are created with a video projector (1280x768, DLP) which projects the image to a semi-transparent screen from behind. Three high-speed, high-resolution FireWire cameras are employed, one of which is prepared with an infrared filter (IR) and a wide-angle lens. It is placed behind the touch screen to receive an IR signal from the screen surface (see Nobuyuki & Jun, 1997). The signal is composed of IR light reflected from users' fingertips on the screen surface, originally emitted by an IR lamp behind the screen, next to the projector. The two other cameras on the top of the screen are used to track the movements of the MagicWands.

3.2 Emotion Tracking

Literature on emotion recognition introduces several techniques to track user emotions (e.g. in Cowie et al., 2001; Picard, 1997). The most salient cues of human emotion are visual and auditory; facial expressions and speech respectively. In PuppetWall, we have started our exploration with in a single modality using auditory cues as the primary input for emotion decoding. This seemed the best option for an application which is intended for multiple, mobile users that can be equipped with headsets. For emotional speech categorization, we utilize a corpus and a feature-based classifier. The enacted corpus (acquired according to specifications described in Seppänen et al., 2003) consisted of long (over 20 sec.) speech extracts with fixed content being voiced with different emotions in a random order. Nine professional actors read the passage ten times. The corpus was used in the training of a feature extractor and a Naïve Bayesian classifier (see Vogt & Andre, 2005, 2006). The classifier allows real-time categorization of four emotion classes represented in the corpus (*neutral, happy, angry, sad*). The result of training the classifier resulted in mediocre off-line recognition rates (55% over the corpus), possible due to very segmenting into short two second samples. In PuppetWall, the output from the classifier was fed into an interpretation layer ("emotional stage" in Figure 1), which transformed the categorical input into a dimensional representation of the emotional state, buffering and smoothing the input.

The use of non-specific corpus was here taken as a starting point. Although the convention is to build a custom corpus, it is generally not a very efficient option and we wanted to investigate possibilities of re-using an existing corpus. The specifications of the corpus seemed promising for our application as it consisted of enacted speech in the target language. From our previous work in the application area (Liikkanen et al., 2008) we knew that there were clear differences in users' manner of speech when they were controlling the characters (being "emerged in the world of play") and when they were addressing each other as their normal selves.

3.3 Emotion Expression

For the current version PuppetWall, we designed a set (characters, objects, background) which covered the Little Red Riding Hood (LRRH) story. LRRH provided a small number of characters and an existing, well-known narrative with distinct phases in the development of emotional tension. Thus we considered it as a good beginning, although the final aim of application development is to enable improvisation in a less constrained environment. The LRRH version thus presents an exploratory step which we used to evaluate the present affective features. The novel

feature in this application is emotion expression achieved through associating detected emotional states to the different, categorically representative forms of the main characters. This relation is depicted in Figure 2.

Figure 2: Four emotional states of the character Wolf from Little Red Riding Hood as visualized with PuppetWall.

4. USER TRIAL

Two professional puppeteers who were not familiar with the operating principles of the application, participated in an evaluation session. The trial (appr. 90 min.) was captured on two DV cameras, one capturing the narrative from the screen, and the other the actors' gestures. The use situation is depicted in Figure 3. Two experimenters were constantly present, one guiding the interactive session, the other providing technical supervision.

Figure 3: Two test users facing PuppetWall, playing out the story of Little Red Riding Hood.

The subjects received minimal instructions verbally. They were informed about the basic interactive features of the system and additionally told about the existence of "interactive features which depend on the emotion recognized from speech". They were requested to play out the story of the Little Red Riding Hood, which they knew by heart from their own productions. They were advised to avoid talking simultaneously as we could not otherwise provide an adequate separation of the speech streams.

4.1 Results

Several interesting qualitative findings were made by observing the session and in the initial analysis of the video recording. During the session the actors created nine different versions of the play, in average little more than four minutes long. They played out the story from the start until the moment where the LLRH is consumed by the Wolf. Users quickly discovered the emotional reactivity of the application. However, the performance of emotion recognition was not satisfactory, the online rates were worse than offline, possibly due to a difference in audio input hardware. As a consequence the behavior of the system seemed erratic to the users.

Finding out that the behavior of the system was not in their full control, the users entered an exploratory phase. They began to experiment with different expressive maneuvers to discover the different forms of the character and how they were related to input. Interestingly, actors used exaggerated aggressive and happy voices together with some distinct words (curses, praises). Also paralinguistic expressions such as crying, laughing, and grunting were observed. The fact that these users adopted these measures without any information regarding how the system worked is notable. Despite the efforts of going through a spectrum of expressions, actors could not reverse engineer the logic of emotion detection. In the final rounds of the session, we saw how the users began to adapt to the limited control situation and started to exploit it as stimuli for improvisation.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a prototype of an interactive, affective application called PuppetWall and documented a preliminary study of affective interaction. The first prototype was built upon an existing corpus which was hypothesized to provide an adequate start for application development. In our evaluation this method turned out to be less successful than what we had hoped for. Because specifying and building a corpus is both a demanding and necessary task for applications of this kind, we will next consider lessons learned from our exploration.

Our investigation suggests that for use with an interactive application, the use context and emotion expression within the application are important considerations right from the start. The corpus should reflect the usage situation, the content and the length of expected expressions. The only way currently to achieve this is seems to be acquisition of a

custom corpus. In our case study, we saw how the users began exaggerate their expressions, after the emotional reactivity they looked for was not otherwise found. However, this has implications regarding potential inputs for the future version of PuppetWall. Instead of mimicking the traditional speech recognition approaches, we could consider a fusion of all available auditory information. Potential improvements can be summarized as follows:

- Paralinguistic information can provide additional cues
- Keyword spotting could be a shortcut to semantics
- Information about the emotional features will affect users' behavior. An important design decision.

Assuming that these suggestions can help to overcome the difficulties in emotion interpretation, we will face the design question of which features of the application should be controlled by this kind of emotional intelligence. In PuppetWall we had chosen a central and well visible functionality, the form of the controlled character. It appeared it was not the kind of feature that should be controlled by an emotion interpreter of mediocre accuracy. Better idea might be to use emotional interpretation to *augment* the user experience (story telling), not to affect the key functional features of the application. For instance, we could imagine using background music, or more abstract character visualizations to represent the emotional states. From the perspective of interaction research and design, there are several unanswered questions remaining and untouched. For instance, how do the affective interaction loops we are looking work? Do we see emotions originating purely from users or do we think that the application can induce emotions? If the latter option is true, what sorts of constraints are there? If we are designing an installation for public spaces and all audiences, could this be a potential ethical issue? How do users affect each other? Are user expressions more important than anything originating from the application? More questions could be added, but in short, there are plenty of questions to be answered by future research on affective interaction.

6. ACKNOWLEDGMENTS

The development of PuppetWall and the preparation of this manuscript were supported by a EU Sixth Framework Program research project CALLAS (ref. 034800).

7. REFERENCES

Camurri, A., Volpe, G., De Poli, G. & Leman, M. (2005). Communicating expressiveness and affect in multimodal interactive systems. *IEEE Multimedia*, 12(1), 43-53.

Cavazza, M., Charles, F., Mead, S.J., Martin, O., Marichal, X. & Nandi, A. (2004). Multimodal acting in mixed reality interactive storytelling. *IEEE Multimedia,* 11(3), 30-39.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine,* 18(1), 32-80.

Liikkanen, L.A., Jacucci, G., Huvio, E., Laitinen, T. & Andre, E. (2008). Exploring Emotions and Multimodality in Digitally Augmented Puppeteering. In *Proceedings of the Advanced Visual Interfaces 2008 (AVI2008)*, Naples, Italy: ACM.

McQuiggan, S.W. & Lester, J.C. (2007). Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies,* 65(4), 348-360.

Nobuyuki, M. & Jun, R. (1997). *HoloWall: designing a finger, hand, body, and object sensitive wall.* Paper presented at the 10th annual ACM symposium on User Interface Software and Technology (UIST).

Paiva, A., Fernandes, M. & Brisson, A. (2006). Children as affective designers - i-shadows development process Technical report from Humaine Workshop on Innovative Approaches for Evaluating Affective Systems. Retrieved from http://www.sics.se/interaction/wp9ws/doc/paiva-wp9ws.pdf.

Paleari, M. & Lisetti, C.L. (2006, October 27). Toward multimodal fusion of affective cues. In *Proceedings of the 1st ACM international workshop on Human-Centered Multimedia*, Santa Barbara, CA, 99-108.

Pantic, M. & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE,* 91(9), 1370-1390.

Picard, R.W. (1997). *Affective Computing*: MIT Press.

Scherer, K.R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication,* 40(1-2), 227-256.

Seppänen, T., Toivanen, J. & Väyrynen, E. (2003). MediaTeam speech corpus: a first large Finnish emotional speech database. In *Proceedings of the Proceedings of XV International Conference of Phonetic Science*, Barcelona, Spain, 2469-2472.

Shugrina, M., Betke, M. & Collomosse, J. (2006). Empathic painting: interactive stylization through observed emotional state. In *Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering (NPAR 2006)*, France: ACM Press, 87-96.

Ververidis, D. & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication,* 48(9), 1162-1181.

Vogt, T. & Andre, E. (2005). Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2005)* Amsterdam, Netherlands: IEEE, 474-477.

Vogt, T. & Andre, E. (2006). Improving Automatic Emotion Recognition from Speech via Gender Differentiation. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2006)*, 1123–1126.