The Workshop Programme

09:00 Welcome (10 min)

New Emotion Corpora (Chair: Laurence Devillers)

09:10 AFFECTIVE LINKS IN A CHILD-ROBOT INTERACTION
Agnes Delaborde, Marie Tahon and Laurence Devillers

09:35 SPEECH IN MINIMAL INVASIVE SURGERY - TOWARDS AN AFFECTIVE LANGUAGE RESOURCE OF REAL-LIFE MEDICAL OPERATIONS
Björn Schuller, Florian Eyben, Salman Can and Hubertus Feussner

10:00 A CORPUS FOR IDENTIFICATION OF SPEAKERS AND THEIR EMOTIONS
Marie Tahon, Agnes Delaborde, Claude Barras and Laurence Devillers

10:25 COFFEE BREAK (20 min)

Linguistics and Emotion (Chair: Anton Batliner)

10:45 INTERFACING WORDNET-AFFECT WITH OCC MODEL OF EMOTIONS
Alessandro Valitutti and Carlo Strapparava

11:10 LINGUISTIC STRUCTURE, NARRATIVE STRUCTURE AND EMOTIONAL INTENSITY
Tibor Pólya and Kata Gábor

11:35 ONLINE TEXTUAL COMMUNICATIONS ANNOTATED WITH GRADES OF EMOTION STRENGTH
Georgios Paltoglou, Mike Thelwall and Kevan Buckley

12:00 LUNCH (90 min)

Multimodal Emotion Corpora (Chair: Roddy Cowie)

13:30 MULTIMODAL RUSSIAN CORPUS (MURCO): STUDYING EMOTIONS
Elena Grishina

13:55 THE ROVERETO EMOTION AND COOPERATION CORPUS
Federica Cavicchio and Massimo Poesio

14:20 THE EMOTIONAL AND COMMUNICATIVE SIGNIFICANCE OF HEAD NODS AND SHAKES IN A NATURALISTIC DATABASE
Roddy Cowie, Hatice Gunes, Gary McKeown, Lena Vaclau-Schneider, Jayne Armstrong and Ellen Douglas-Cowie

Annotation and Similarity of Emotion Corpora (Chair: Ellen Douglas-Cowie)

14:55 ANNOTATION OF THE AFFECTIVE INTERACTION IN REAL-LIFE DIALOGS COLLECTED IN A CALL CENTER
Christophe Vaudable, Nicolas Rollet and Laurence Devillers

15:20 PRESENTING THE VENEC CORPUS: DEVELOPMENT OF A CROSS-CULTURAL CORPUS OF VOCAL EMOTION EXPRESSIONS AND A NOVEL METHOD OF ANNOTATING EMOTION APPRAISALS
Petri Laukka, Hillary Anger Elfenbein, Wanda Chui, Nutankumar S. Thingujam, Frederick K. Iraki, Thomas Rockstuhl and Jean Althoff

15:45 TOWARDS MEASURING SIMILARITY BETWEEN EMOTIONAL CORPORA
Mátyás Brendel, Riccardo Zaccarelli, Björn Schuller and Laurence Devillers

16:10 COFFEE BREAK (40 min)

Extended and Multiple Emotion Corpora (Chair: Björn Schuller)

16:50 INDUCED DISGUST, HAPPINESS AND SURPRISE: AN ADDITION TO THE MMI FACIAL EXPRESSION DATABASE
Michel F. Valstar and Maja Pantic

17:15 COMPLEMENTING DATASETS FOR RECOGNITION AND ANALYSIS OF AFFECT IN SPEECH
Tal Sobol-Shikler

17:40 CROSS-CORPUS CLASSIFICATION OF REALISTIC EMOTIONS - SOME PILOT EXPERIMENTS
Florian Eyben, Anton Batliner, Björn Schuller, Dino Seppi and Stefan Steidl

18:05 Closing Discussion (25 min)

18.30 End of workshop (followed by an informal dinner)

Workshop Organisers

Laurence Devillers / Björn Schuller
LIMSI-CNRS, France

Roddy Cowie / Ellen Douglas-Cowie
Queen's University, UK

Anton Batliner
Universität Erlangen-Nürnberg, Germany

Workshop Programme Committee

Laurence Devillers / Björn Schuller
LIMSI-CNRS, France

Roddy Cowie / Ellen Douglas-Cowie
Queen's University, UK

Anton Batliner
Universität Erlangen-Nürnberg, Germany

Table of Contents

Annotation and Similarity of Emotion Corpora

Extended and Multiple Emotion Corpora

Author Index

Introduction

Recognition of emotion in speech has recently matured to one of the key disciplines in speech analysis serving next generation human-machine and –robot communication and media retrieval systems. However, compared to automatic speech and speaker recognition, where several hours of speech of a multitude of speakers in a great variety of different languages are available, sparseness of resources has accompanied emotion research to the present day: genuine emotion is hard to collect, ambiguous to annotate, and tricky to distribute due to privacy preservation.

The few available corpora suffer from a number of issues owing to the peculiarity of this young field: as in no related task, different forms of modelling reaching from discrete over complex to continuous emotions exist, and ground truth is never solid due to the often highly different perception of the mostly very few annotators. Given by the data sparseness – most widely used corpora feature below 30 min of speech – cross-validation without strict test, development, and train partitions, and without strict separation of speakers throughout partitioning are the predominant evaluation strategy, which is obviously sub-optimal. Acting of emotions was often seen as a solution to the desperate need for data, which often resulted in further restrictions such as little variation of spoken content or few speakers. As a result, many interesting potentially progressing ideas cannot be addressed, as clustering of speakers or the influence of languages, cultures, speaker health state, etc..

Previous LREC workshops on Corpora for research on Emotion and Affect (at LREC 2006 and 2008) have helped to consolidate the field, and in particular there is now growing experience of not only building databases but also using them to build systems (for both synthesis and detection). This workshop aims to continue the process, and lays particular emphasis on showing how databases can be or have been used for system building.

Overall, the topics of this workshop in the area of corpora for research on emotion and affect include, but are not limited to:

- Novel corpora of affective speech in audio and multimodal data – in particular with high number of speakers and high diversity (language, age, speaking style, health state, etc.)
- Case studies of the way databases have been or can be used for system building
- Measures for quantitative corpus quality assessment
- Standardisation of corpora and labels for cross-corpus experimentation
- Mixture of emotions (i.e. complex or blended emotions)
- Real-life applications
- Long-term recordings for intra-speaker variation assessment
- Rich and novel annotations and annotation types
- Communications on testing protocols
- Evaluations on novel or multiple corpora

The organizing committee

Laurence Devillers / Björn Schuller
Spoken Language Processing group
LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
devil@limsi.fr / schuller@limsi.fr

Roddy Cowie
School of Psychology
Ellen Douglas-Cowie
Dean of Arts, Humanities and Social Sciences
Queen's University
Belfast BT7 1NN, UK
r.cowie@qub.ac.uk / e.douglas-Cowie@qub.ac.uk

Anton Batliner
Lehrstuhl fuer Mustererkennung (Informatik 5)
Universität Erlangen-Nürnberg
Martensstrasse 3, D-91058 Erlangen, Germany
batliner@informatik.uni-erlangen.de

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

Affective Links in a Child-Robot Interaction

Agnes Delaborde, Marie Tahon, Laurence Devillers

LIMSI-CNRS
BP 133, 91 403 Orsay cedex, France
(agnes.delaborde|marie.tahon|laurence.devillers)@limsi.fr

Abstract

This article is an introduction to a novel corpus featuring children playing games in twos with the humanoid robot Nao by Aldebaran Robotics. The robot's behaviours are remotely controlled. In the course of these games, children's emotional reactions are triggered through specific strategies. This is also a case study of the way a corpus of paralinguistic clues can provide a groundwork for bringing context to an emotional system through emotion detection. We also question the clues about affective links that can be infered from the corpus annotations, and the emotions expressed by children according to the robot's behaviors.

1. Introduction

The NAO-Children corpus features children playing in twos with a remotely operated humanoid robot, Nao[1]. Their interactions are controlled by strategies meant to elicitate emotional reactions, in the same vein as (Batliner et al., 2004), the difference being that children play with the robot in groups of two, and that Nao answers back. This allows a dialogue interaction to take place, between the child and Nao, and between children.

The corpus is being collected in the course of the Project ROMEO[2]. The aim of the project is to design a robotic companion which can play different roles; we focused, in this paper, on the role of the robot as a game companion. We worked in our first experiments with the Nao prototype. In its final role, the robot has to be able to supervise a game, while also being sensitive to the emotions of the children. It should for example be able to detect through the expressed emotions if a child is sad. It also has to behave in a way such as to entertain the children, and maintain their desire to play with it.

The NAO-Children corpus will allow to train models on real-time emotion detection in children's voice, in a Human-Robot Interaction context. This will also allow studies on speaker identification robust to emotional speech, and emotional interaction models of children at play with a social robot.

In a first part, we look into the contextual information which is required in an emotional detection, and the way emotion detection would bring more contextual clues. We present in a second part the NAO-Children corpus, its acquisition and its annotations. We conclude with our first remarks about the children's reactions during the games, and the affective links that can be observed in this corpus.

2. Emotional System

In the framework of the ROMEO Project will be designed an emotional system for the robot. It will allow it to be able to adapt its behaviour according to its own emotional state, and will be sensitive to the user's emotional state as well.

The robot will evolve in real-life conditions, then face a rich contextual environment which needs to be processed. Models of emotional system used in storytelling tasks are complete, for they take into account an almost full context of interaction: be it the non-human entity's past and memory (Ochs et al., 2009; Bickmore et al., 2009), its relation to the other humans or non-human entities around it (Rousseau and Hayes-Roth, 1998; Kaplan and Hafner, 2004; Michalowski et al., 2006; Sidner et al., 2004), or a modeling of the events happening in the agent's surroundings (Ochs et al., 2009), etc.

However, emotions in real-life conditions are complex, and factors responsible for the emergence of an emotional manifestation are intricate (Scherer, 2003). The contextual information cannot be as easily scenarised as in a storytelling system, and new clues about context have to be added to real-life systems. Detecting automatically the emotion from the human voice is in itself a real challenge: without any clues about the context of emergence, the events or emotional dispositions of the speaker, the detection only relies on prosodic features. We will try to look into contextual clues (such as profile of the speaker, age, etc.) which we can detect from the audio signal and bring to an emotional system.

An audio corpus annotated with interaction and emotional information will provide a basis. To design such a corpus, we first need it to be task-related, i. e. it has to feature children having gamin interactions with a robot. We also need the expressions of emotions we gather to be spontaneous and numerous enough. As we need a scientific control over the emotions being expressed by the children, we elicitate them through some specific behaviours in the robot.

3. NAO-Children Corpus

Designing affective interactive systems needs to rely on an experimental grounding (Picard, 1997). However, cost, or privacy, are disuassive in the creation of an emotional corpus based on real-life or realistic data (Douglas-Cowie et al., 2003).

In the NAO-Children corpus (Delaborde et al., 2009), two children by session are recorded as they play with the robot. A game master supervises the game, gives the question cards, and encourages the children to interact with the

[1] http://www.aldebaran-robotics.com

[2] Cap digital French national project founded by FUI6, http://www.projetromeo.com

robot. In order to reinforce the emotional reactions of the children, only friends or sisters and brothers are recorded. The robot in this context is a player, and also tries to answer the questions. In order to trigger emotional reactions in the children, it acts in an unconform way from times to times.

3.1. Corpus Characteristics

So far, ten French children (five girls, five boys), aged between eight and thirteen years, have been recorded with high-quality lapel-microphones. Recordings amount to two hours of French emotional speech. Data are being segmented and annotated, and six children's recordings have been annotated yet. We plan to carry on with the annotation, and meanwhile gather some more recordings, so as to record a sum total of around fifty different children.

3.1.1. Objectives of the Corpus

The corpus aims at gathering recordings of children playing in a family setting (brother and sisters, or friends). It provides acted, induced and spontaneous emotional audio data. These data will be a groundwork for studies on emotion detection in Human-Robot Interaction, speaker identification robust to emotional speech, and emotional interaction models of children at play with a robot.

3.1.2. Protocole

Children were offered to play games with the robot.
1) Questions-Answers game: each player reads by turns a question written on a card, and the two others try and guess the answer.
2) Songs game: each human player has to hum the song which title is written on a card, until the robot recognizes the song.
3) Emotions game: each human player acts an emotion (Fear, Joy, Anger, Sadness), and the robot says aloud what it detected. The child is offered to act again, until the robot recognizes it correctly.
The robot is remotely operated by a Wizard of Oz, who loads predetermined behaviors meant to have the children react. In the course of the game, the robot plays several roles. It will be an attentive game player and quietly answer or help to answer the questions. But, according to our elicitation strategies, it will also go off when the children do not expect it, refuse to help a child, favor one child rather than the other, or mix up the rules. These strategies are detailed in Table 1, which represents how many times we try to elicitate an emotion in the children through the robot, in the course of the game. Some strategies are directed to both children (and thus are counted for both).
In the course of the Emotions and Songs games, elicitation strategies are used alternatively, according to the Wizard-of-Oz experimenter's perception of the mental state of the child : if a child seems to be good-willing and relaxed, the robot will not detect correctly his or her song/acted emotion (it will detect an emotion of the opposite valence for example), and it will detect nearly instantly the other child's emotion/song. This strategy is applied up to a certain threshold : when the first child seems to get tired of playing, the detection works correctly. A shy or nervous child will succeed faster.

Strategy	Directed to Child A	Directed to Child B
Motivates, prompts to interaction	1	1
Congratulates	1	2
Persists in not understanding the child	2	2
Does not understand the game	2	2
Generates competition	2	0
Crashes	1	1

Table 1: Strategies used during the Question-Answer game, and the number of times they are directed to children

3.1.3. Annotation Scheme

On each child's track, we define segment boundaries. A segment is emotionally homogenous, i. e. the emotion is considered as being the same and of a constant intensity along the segment (Devillers et al., 2006, Devillers and Martin, 2008).
Each segment is at the present time annotated by two expert labelers, using the Transcriber annotation tool[3].

- Affective state label

So as to describe the complexity of the expressed emotions, three affective state labels (cf. Table 2) are used to describe each segment. The first label describes the most salient affective state that is perceived by the annotator. The two others define more precisely the first label. Every permutation is possible: positive labels can be mixed with negatives; there can also be no perceived emotions at all.

- Classical Dimensions
  - Valence

    Does the speaker feel a positive or a negative sensation? *positive*, *negative*, *ambiguous: either positive or negative*, *positive and negative*, *valence non decidable*.

  - Intensity

    The strength of the expressed emotion. Scale of 1 to 5, from *very weak* to *very strong*.

  - Activation

    How many phonatory means are involved in the expression of the emotion (sound level, trembling voice, over-articulation, etc.)? Scale of 1 to 5, from *very few* to *a lot*.

  - Control

    Does the speaker control, contain his or her emotional reaction? Scale of 1 to 5, from *not at all* to *completely*.

[3] http://trans.sourceforge.net/en/presentation.php

Affective state's Category	Annotation value
POSITIVE	Joy Amusement Satisfaction Motherese Positive
ANGER	Anger Irritation
SADNESS	Sadness Disappointment
FEAR	Fear Anxiety Stress Embarrassment
NEUTRAL	Neutral
OTHERS	Surprise Interest Empathy Compassion Irony Scorn Negative Boredom Excitation Provocation Overbid

Table 2: Affective state labels

- Mental States (Zara et al., 2007; Baron-Cohen et al., 2000)

By observing the speaker talking, his or her emotional reaction, what can we infer about his or her thoughts, desires or intentions? E. g. *to be sure*, *to doubt*, *to agree*, etc.

- Trigger Event

What kind of event triggered the child's emotional reaction? If the trigger comes from the other child, what type of communication act? E. g. *encourages*, *laughs at*, *laughs with*, *explains*.

If the trigger comes from the robot, what type of elicitation strategy? E. g. *asks for attention*, *encourages*, *inappropriately goes off*, *never recognizes the child emotion/song*.

If the trigger comes from the game master, what type of communication act? E. g. *child control - security*, *explains the rules*, *the game master reinforces an elicitation strategy that failed*.

- Spontaneous/Acting

This value is a flag that will allow, in subsequent corpus processings, to spot whether the speaker was at that moment acting an emotion in the course of the game, or if he or she was reacting spontaneously to an event. Values are *Spontaneous* or *acted*.

4. Affective Links

The NAO-Children corpus gathers annotated emotional data, which allow us to test the affective interaction strategies applied through the robot. The interest of this data collection is twofold: we observe children interacting with each other, and with the robot.

The corpus provides a basis for observing the children's reactions according to the different robot's behaviors. For example, when the robot makes mistakes, young girls tend to mother the robot and explain it patiently (involvment in the interaction). On the other hand, teenage boys seem more inclined to condescend to it, and make use of irony (disengagement). Both girls and boys sometimes simply laugh at the robot's obvious mistakes. The aim being to bring pleasure and comfort to the interaction, what strategies should the robot apply to please the child? We can study in the corpus annotations whether children are pleased with positive attitudes from the robot, and on the contrary unpleased by negative ones, or if there are more intricate patterns.

From the first annotations, we can draw some verifications about the settings of the experiment. We observe the global feeling labelers had about the children recordings' valence, when the latter were reacting spontaneously. On four children (three boys and one girl), they considered on average that for a half of our first study corpus, children expressed positive emotions during the recording sessions. The proportion of negative perceived emotions vary, though: one labeler considered that for 30% of the segments the children expressed un-valenced emotions, and for 13% their emotions were negative (the remaining consisting of ambigous or positive-negative emotions). The other labeler felt that 16% of the expressed emotions were negative, 15% consisted of neutral emotions, and 15% of ambigous valence emotions. This first approach of our data, if this is to be confirmed by the future annotations, tends to show that the settings, on the whole and in spite of the experimental setting, expressed a good proportion of valenced emotions during their interaction with the robot.

When a strategy applied through the robot fails (for example the robot goes off in the middle of the game, when the children thus expect it less), the game master has to draw the attention of the children towards this fact. If we question the influence of the presence of the game master in the settings, i. e. if he does not prevents the child-robot interaction to take place, we notice so far that when the game master reinforces a strategy, children speak directly to the robot, so far, in 70% of the cases.

Besides, the presence of two children interacting with the robot complexifies the interaction. Is a younger brother's reaction altered by the presence of his older brother? For example, a younger brother expresses irritation while his brother tries to explain him how to win during the Emotions game. Will this kind of interactions triggers some different types of reaction in the child, which must be taken into account?

Studies on social interaction between human and non-human entities, and its influence on emotion expressions (Ochs et al., 2009, Bickmore and Cassell, 2005, Isbister, 2006) could allow us to determine some tendency of the speaker's emotional dispositions. By allowing an adaptive

and dynamic processing of the affective links between the human and the robot, emotion detection can bring valuable indications to maintain the interaction.

5. Conclusion

The NAO-Children corpus is composed of French emotional speech recorded from children from eight to thirteen, annotated with emotional and interactional labels. The corpus presents children interacting in a game context with a humanoid robot, which applies strategies meant to elicitate emotional reactions in them. This corpus allows to study new paralinguistic clues which can bring more context to an emotional system.

The recordings we gathered so far will allow us to test our settings and strategies. We expect to record children up to a sum total of fifty children, so as to get relevant information about the interactions between children and the robot, and to get a corpus large enough for emotion detection training. A future application could be an interactive stories game, during which the users could take part in the construction of the story. Their emotional reactions would be taken into account, so as to modify the course of the game and increase their gaming pleasure.

6. References

S. Baron-Cohen, H. Tager-Flusberg, and D. J. Cohen. 2000. *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*. Oxford University Press.

A. Batliner, C. Hacker, S. Steidl, E. Nth, S. D'Arcy, M. Russel, and M. Wong. 2004. You stupid tin box - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proc. of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

T. Bickmore and J. Cassell. 2005. Social Dialogue with Embodied Conversational Agents. *Advances in Natural Multimodal Dialogue Systems*, 30:25–54.

T. Bickmore, D. Schulman, and L. Yin. 2009. Engagement vs. Deceit: Virtual Humans with Human Autobiographie. In *Proc. of Intelligent Virtual Agents*, Amsterdam, The Netherlands.

A. Delaborde, M. Tahon, C. Barras, and L. Devillers. 2009. A Wizard-of-Oz Game for Collecting Emotional Audio Data in a Children-Robot Interaction. In *Proc. of the International Workshop on Affective-aware Virtual Agents and Social Robots, ICMI-MLMI*, Boston, USA.

L. Devillers and J.-C. Martin. 2008. Coding Emotional Events in Audiovisual Corpora. In *Proc. of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

L. Devillers, R. Cowie, J.-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie. 2006. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. In *Proc. of the International Conference on Language Resources and Evaluation*, Genoa, Italy.

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

K. Isbister. 2006. *Better Game Characters by Design: A Psychological Approach*. Morgan Kaufmann Publishers Inc.

F. Kaplan and V. V. Hafner. 2004. The Challenges of Joint Attention. In *Proc. of the Fourth International Workshop on Epigenetic Robotics*, Genoa, Italy.

M. P. Michalowski, S. Sabanovic, and R. Simmons. 2006. A Spatial Model of Engagement for a Social Robot. In *Proc. of the 9th IEEE International Workshop on Advanced Motion Control*.

M. Ochs, N. Sabouret, and V. Corruble. 2009. Simulation of the Dynamics of Non-Player Characters Emotions and Social Relations in Games. In *Transactions on Computational Intelligence and AI in Games*.

R. W. Picard. 1997. *Affective Computing*. MIT Press.

D. Rousseau and B. Hayes-Roth. 1998. A social-psychological model for synthetic actors. In *Proc. of the International Conference on Autonomous Agents*, pages 165–172, Minneapolis, MN, USA. ACM.

K. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256.

C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh. 2004. Where to Look: A Study of Human-Robot Engagement. In *Proc. of the 9th international conference on Intelligent user interfaces*, Funchal, Madeira, Portugal.

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 *Proc. of the second International Conference on Affective Computing and Intelligent Interaction*, Lisbon, Portugal.

Speech in Minimal Invasive Surgery – Towards an Affective Language Resource of Real-life Medical Operations

Björn Schuller[1], **Florian Eyben**[1], **Salman Can**[2], **Hubertus Feussner**[2,3]

[1] Institute for Human-Machine Communication, Technische Universität München, Germany
[2] Research Group MITI, Klinikum rechts der Isar der Technische Universität München, Germany
[3] Klinik und Poliklinik des Klinikums rechts der Isar der Technische Universität München, Germany
schuller@tum.de

Abstract

There is a clear desire to collect language resources of utmost realism with respect to spontaneity of speech and naturalness of emotion. This goal is difficult to obtain as a setting is needed that provides sufficient 'emotional moments' while these shall not be disturbed by the speakers' awareness of recording. An obvious setting seems to be collecting surgeon's speech during real-life operations: considering the responsibility of patients' health and life, emotion can be assumed to be present and at the same time natural – there simply is not the time of wasting thoughts on the fact that one is being recorded.

1. Introduction

There is an ever present demand for realistic emotion recorded in a natural context and environment due to a number of prohibitive factors as fore mostly privacy of recorded subjects, their awareness of the recording situation, presence of noises and an ever ambiguous ground truth (Devillers et al., 2005; Douglas-Cowie et al., 2003; Schuller et al., 2009c; Ververidis and Kotropoulos, 2003; Zeng et al., 2009).

A number of efforts have been undertaken leading to some of the most popular databases of emotional speech with 'realistic' emotions. Yet, often these were recorded in experimental settings rather than in true every-day life situations, as in Wizard-of-Oz experiments (e. g. the SmartKom database (Steininger et al., 2002) of user interaction with a smart information kiosk, the FAU Aibo database (Steidl, 2009) of Child-Robot Interaction, or the "Sensitive Artificial Listener" (SAL) database (Douglas-Cowie et al., 2007) of Human-Chatbot conversation. The "Audio-Visual Interest Corpus" (AVIC) (Schuller et al., 2009b) is an example of human-to-human conversational speech (a presenter and a subject experiencing different levels of interest throughout a product presentation) not recorded in a simulation, but still in an experimental framework: the subjects were invited without their own original interest to watch the presentation. The "Vera-Am-Mittag" (VAM) corpus is a popular representative of data from broadcasts – here a reality TV show – yet, appearing in a TV show is hardly to be considered as every-day life situation for the target subjects, and in addition it is not fully known to which degree such a show might be scripted in advance. Finally, the "Speech Under Simulated and Actual Stress" (SUSAS) set (Hansen and Bou-Ghazale, 1997) is a well-known example of subjects recorded in their actual work situation – steering a military chopper – yet, still a pre-defined protocol was followed, as these had to speak pre-defined one-word commands (a 20 items vocabulary) in distinct situations as starting, landing, etc..

Considering the above respects, we decided to establish a language resource of surgeon speech during real-life operations in their usual operation room: this is their work routine, emotion is present in a 'less than 90 % neutral' distribution and these can be assumed to be sufficiently natural, as the surgeons arguably simply do not have time to waste a thought on being recorded when lives are at stake, and in addition repeated sessions rather than a single dialog (as e. g. often the case for call centre data, e. g. (Burkhardt et al., 2005)) can be collected: our "Speech in Minimal Invasive Surgery" (SIMIS) database consists of speech recorded during medical operations and has so far mostly been used for improvement of Automatic Speech Recognition (ASR) for control of assisting surgery robots (Schuller et al., 2008; Schuller et al., 2009a; Munoz et al., 2000; Allaf et al., 1998; Hurteau et al., 1994). However, in this paper we introduce details on its labelling with respect to affect to provide a further language resource of affective speech for research purposes.

It consists of 29 live recordings of surgery which we will call Set A in the ongoing. This whole Set A has been textually transcribed and annotated with five affective states by a single male labeller '$L1$'. During the work reported in this paper, 6 new life recordings - referred to as Set B - were added to the original SIMIS Database, which thus now consists of 35 live recordings which we will call Set $A + B$. They were all textually transcribed and annotated within the same set of emotions by one female labeller '$L2$' stemming from the same age group (20–30 years) as the labeller $L1$.

The desire behind this effort is to establish a more reliable ground truth of emotion annotation and the fact that prior to this study the SIMIS database comprised recordings of male German surgeons only. To draw more significant conclusions on the subject of speaker-independent emotion-recognition, it was desirable to include female speakers as well as non-native speakers. In the present version, both, a female person and a Turkish surgeon have been recorded.

In this paper we will provide details on the recordings (Section 2.), segmentation (Section 3.), annotation (Section 4.), and linguistic analysis with respect to emotion classes.

Figure 1: The operating room of the Clinic r. d. Isar where all surgeries were recorded

Figure 2: The operating room of the Clinic r. d. Isar during a surgery

2. Recordings

To work with real life emotions is crucial in our case, because we want to create a reliable database, which can be used in real-life-situations. Moreover, it is important to record a variety of speakers to achieve a reliable speaker independent emotion-recognition.

The Clinic *Rechts der Isar* (abbreviated r. d. Isar in the ongoing) of TUM in Munich, Germany (shown in Figure 1) was selected for the recordings.

Usually during an operation there are 6 to 10 people present in the operating room. The surgeon, 2 to 3 assistants and 3 to 6 nurses conditioned by the complexity of the operation and the experience of the surgeon. The operating room during surgery is shown in Figure 2.

During an operation there is a great amount of background noise: several assisting machines run during the operation, telephones are ringing, the nurses talk to each other and sometimes a radio is playing (Schuller et al., 2009a). Additionally, the entire room is tiled to fulfil the hygiene-rules, and so there are diffuse acoustical reflections that may result in increased background-noise level.

For speech capturing we decided for the AKG C 444 L wireless headset. This device possesses a cardioid pattern. The sagger of the microphone is optimised for speech in the near field. The low-frequency transmission is reduced, so the typical near field-effect of the velocity microphone is shaken out. Because of that the speak-distance of $2\,cm$ between 80 and $5\,kHz$ is nearly linear. Greater speak distances require a greater compensation of low-frequencies.

An AKG PT 40 sender transmitted the data along a quartz stabilised carrier frequency in the UHF domain, and an AKG SR 40 received it. The data was stored with $16\,bit$ per sample and a sample rate of $16\,kHz$ to hard disk drive.

Table 1: Operation types, number of recordings, and average duration.

Operation	#	**Average Duration** *[min.]*
Set A		
Gall	17	57
Fundoplicatio	6	103
Sigma Wedge	6	108
Total	**29**	**77**
Set B		
Gall	3	43
Umbilical hernia	1	84
Vakusil	1	23
Thyroid	1	119
Total	**6**	**59**
Set $A + B$		
Gall	20	55
Fundoplicatio	6	103
Sigma Wedge	6	108
Umbilical hernia	1	84
Vakusil	1	23
Thyroid	1	119
Total	**35**	**74:04**

2.1. Set A

Prior to this study the SIMIS database consisted of 29 live recordings with a total duration of $2\,240 : 59\,min$ (i. e. over $37\,h$), and a total speech time of $350 : 01\,min$ (i. e. nearly $6\,h$ – for details on segmentation refer to 3.). The operations were recorded from seven surgeons while three different surgery-types have been recorded. The specific length of an operation is based on its complexity - the average duration is shown in Table 1. The surgeons were solely male and native Germans (cf. Table 2).

2.2. Set B

During the study presented in this paper 6 new live recordings with a duration of $356\,min$ (i. e. nearly $6\,h$) were added. Different surgeries with an average duration of $59\,min$ have been recorded (cf. Table 1). The new recordings include five male and one female speakers, one of them with Turkish as his mother language, the others as before

Table 2: Details of the recorded surgeons. Abbreviations: TRT: Total Recording Time (in minutes); # rec.: number of recorded operation sessions.

ID	Native	Gender	Age [years]	# rec.	TRT *[min.]*
			Set A		
S 00	German	male	54	20	1 289
S 01	German	male	46	3	421
S 02	German	male	38	2	211
S 03	German	male	36	1	126
S 04	German	male	35	1	72
S 05	German	male	33	1	67
S 06	German	male	29	1	56
			Set B		
S 00	German	male	54	1	30
S 01	German	male	46	2	137
S 06	German	male	29	1	55
S 07	German	female	34	1	84
S 08	Turkish	male	39	1	23
			Set $A + B$		
S 00	German	male	54	21	1 319
S 01	German	male	46	3	558
S 02	German	male	38	2	211
S 03	German	male	36	1	126
S 04	German	male	35	1	72
S 05	German	male	33	1	67
S 06	German	male	29	1	111
S 07	German	female	34	1	84
S 08	Turkish	male	39	1	23

native German speakers (cf. Table 2).

2.3. Set $A + B$

The extended SIMIS database (Set $A + B$) consists of 35 recordings of 9 surgeons during 6 surgery-types with an average duration of $74\,min$ shown in table 1 and has a duration of $2\,597\,min$ (i.e. over $43\,h$).
Table 2 gives an overview of the surgeons and the amount of data recorded from each of them.

3. Segmentation

As mentioned above, these recordings contain not only speech, but considerable amounts of background noise, too. The most common noise types during surgery are: standard background noise, instrument click noise, background talk, stressed breath or cough from the surgeon (a detailed analysis of the distribution of noise types concerning Set A is found in (Schuller et al., 2009a)). To generate a database for speech-based emotion-recognition, we need to extract turns that contain speech of the recorded surgeon. To test and train emotion recognition systems in future studies, automatic segmentation and silence removal were thus performed.
For this paper the Set $A+B$ recordings were segmented by applying a root-mean-square energy threshold of 0.01 (the samples were normalised to the range $[-1; +1]$. Thereby a minimum turn length of $0.16\,sec$ and a minimum silence length of $0.3\,sec$ was enforced. Several hundred speech turns with an average duration of about $2\,sec$ were obtained from each recording session. The number of segments per recording reached from 86 to 913, depending on the amount of speech and its scattering, while the total speech time took from about $5\,min$ to $39\,min$ (cf. Table 3). For the 35 operations in the SIMIS database a total of 11 077 speech turns were attained, as also shown in Table 3.

Table 3: Segmentation Results. Abbreviations: TRT: Total Recorded Time, TST: Total Speech Time, ST: Speech Turns.

Operation	#	TRT *[min.]*	TST *[min.]*	ST #
		Set A		
Gall	17	975	162	3 952
Fundoplicatio	6	616	72	2 245
Sigma Wedge	6	650	90	2 676
Total	**29**	**2 241**	**324**	**8 873**
		Set B		
Gall	3	130	24	867
Umbilical hernia	1	84	10	292
Vakusil	1	23	5	132
Thyroid	1	119	32	913
Total	**6**	**356**	**71**	**2 204**
		Set $A + B$		
Gall	20	1 105	186	4 819
Fundoplicatio	6	616	72	2 245
Sigma Wedge	6	650	90	2 676
Umbilical hernia	1	84	10	292
Vakusil	1	23	5	132
Thyroid	1	119	32	913
Total	**35**	**2 597**	**395**	**11 077**

4. Annotation

During the annotation the turns were assigned manually one of the following five classes of emotions: *angry* (ANG), *confused* (CON), *happy* (HAP), *impatient* (IMP) and *neutral* (NEU). The choice of these classes bases on a first inspection of the data by an expert. The content of Set A has then been labelled by two labellers, $L1$ and $L2$. At any stage it was required that no turns were skipped or omitted, since the envisioned applications demand that every speech turn has to be dealt with.
If we a look at the emotion distribution by percentage of the emotion-classes annotated by the first annotator and compare it to the percentage of the second annotator as depicted in Table 4, we can see clear deviations, though *neutral* is by far the most common emotion. Besides, labeller $L1$ labelled the non-neutral emotion classes (*angry*, *confused*, *happy*, *impatient*) relatively in balance ($\pm$ max. 3.3 %). Moreover, labeller $L2$ chose the emotion class *impatient* to be the second most frequent while *happy*, *angry* and *con-*

Table 4: Distribution of the emotions (ANG, CON, HAP, IMP, NEU) in percent for the diverse sets (A and B) and per annotator ($L1$ and $L2$).

Set	**Labeller**	ANG	CON	HAP	IMP	NEU
A	$L1$	6.4	9.8	8.0	8.4	67.4
A	$L2$	2.7	2.8	3.7	13.4	77.1
B	$L2$	1.5	5.2	4.3	17.3	64.1

Table 5: Confusions among annotators in Set A ($L1$ to the right). κ=0.56.

#	ANG	CON	HAP	IMP	NEU
ANGER	**151**	8	8	196	167
CONFUSED	6	**142**	11	44	598
HAPPY	8	14	**132**	46	626
IMPATIENT	25	7	17	**335**	388
NEUTRAL	56	105	142	682	**4 959**

fused have been annotated at relatively equal frequency ($\pm$ max. 0.8 %). The agreement of both annotators will be investigated by Cohen's Kappa (no straight forward ordinal relation exists among classes. Naturally, one could be established, though, e. g. by arousal or valence dimensions).

Inter-Labeller-Agreement

To measure the agreement between the labellers we will use Cohen's Kappa (Cohen, 1968). The measurement of the agreement follows a confusion matrix among labellers which is shown in figure 5.

5 579 tracks out of Set A have been annotated in agreement. In 3 291 cases confusion occurs. The allotment of agreement of the labellers p_o is compared by the *accidental*-agreement p_e.

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \quad (1)$$

The difference between p_0 and p_e represents the contingent of cases in which *accidental*-agreement occurs. It is normalised by $1 - p_e$ that is expected by chance. The measured Kappa value for this setting resembles 0.56, which can be considered as good, given the highly subjective nature of spontaneous emotions.

The Kappa value can be improved by combining some of the 5 classes to reduce the problem to a three class valence problem. We thereby map angry and impatient to the class *negative* (NEG), confused and neutral to the class *neutral* (NEU), and happy to the class *positive* (POS). A kappa value of 0.62 is obtained for this reduced 3 class set with the according confusions shown in 6.

5. Linguistic statistics

For all speech turns the spoken text is transcribed by $L1$. For set A, we analysed the mean number of words per turn (l_μ), the standard deviation of l_μ (σ_μ), and the size of the active vocabulary (n_{voc}) for each of the five emotion classes. The active vocabulary size is thereby the number of unique words which occur in all turns belonging to one emotion class. We also report the number of vocabulary items which are unique to one emotion class, i. e. occur only in turns with the respective emotion label (this number is referred to as n_{emo}). The results are shown in table 7 for emotion classes as assigned by labeller $L1$ and in table 8 for emotion classes as assigned by labeller $L2$.

Table 6: Confusions among annotators in Set A ($L1$ to the right) with the emotions reduced to 3 valence-motivated classes (explanation in the text). κ=0.62.

#	NEG	NEU	POS
NEGATIVE	**707**	570	25
NEUTRAL	788	**5 804**	153
POSITIVE	54	640	**132**

Table 7: Linguistic Statistics (Set A), labeller $L1$. Number of turns per emotion class (N_t), average turn length in words (l_μ), standard deviation of turn length in words (l_σ), size of active vocabulary in words (n_{voc}) and the number of words specific to a single emotion (n_{emo}).

Emotion	N_t	l_μ	l_σ	n_{voc}	n_{emo}
ANGER	530	5.6	4.0	836	203
CONFUSED	801	4.7	2.9	897	203
HAPPY	826	5.1	4.1	1 119	364
IMPATIENT	772	4.2	3.4	677	106
NEUTRAL	5 944	5.5	3.8	4 140	2 861

No clear tendency for each class can be deduced from tables 7 and 8, with the exception of the class *Impatient*. Impatient turns marked by labeller $L1$ are clearly shorter than neutral turns (5.3/5.4 words) and the average length of all turns (2.2 seconds with 5.3 words, and a standard deviation of 4.0 words). Second shortest for labeller $L1$ are confused turns. For confused and impatient turns assigned by $L1$ a smaller turn length standard deviation is notable. This may be an indication that $L1$ did take turn length into account when assigning an emotion. The vocabulary used within impatient turns ($L1$) is also notably smaller than for the other three emotions (besides neutral). This indicates that impatient turns may be short turns composed of fewer, simpler command words, or words are repeated by surgeons when they are impatient. For $L2$ the situation seems more balanced, and the conclusions drawn

Table 8: Linguistic Statistics (Set A), labeller $L2$. Number of turns per emotion class (N_t), average turn length in words (l_μ), standard deviation of turn length in words (l_σ), size of active vocabulary in words (n_{voc}) and the number of words specific to a single emotion (n_{emo}).

Emotion	N_t	l_μ	l_σ	n_{voc}	n_{emo}
ANGER	246	4.8	3.6	482	100
CONFUSED	274	5.0	3.4	479	88
HAPPY	310	5.5	4.3	567	119
IMPATIENT	1 303	4.8	3.6	1 245	335
NEUTRAL	6 740	5.4	4.0	4 410	3 246

from the $L1$ statistics must be carefully analysed. Impatient (now along with angry) turns remain shorter than turns of the other four classes.

The total vocabulary size of the set A is 5 078 words. Only 4 140 ($L1$) or 4 410 ($L2$) words of these 5 078 words are found in neutral turns. Thus, we can conclude that high percentage of words used in emotionally coloured turns is not used in neutral turns, and therefore characterises emotionally coloured turns. This is further supported by the numbers in the last column (n_{emo}) of tables 7 and 8, which state that roughly one fifth of the vocabulary used in emotionally coloured turns is used only in turns of the respective emotion and does not appear in turns assigned to any other emotion class. For future classification experiments it may thus be very beneficial to include linguistic information as features.

6. Conclusion

Our study again proves the challenge of dealing with emotions in a real-life scenario: moderate inter labeller agreement is observed which is typical for such spontaneous and naturalistic emotion classification tasks, e. g. (Schuller et al., 2009c). A slight correlation between turn-length (measured in words) and impatient or angry turns has been found. Due to the fact that approximately one fifth of the vocabulary used in emotionally coloured turns of one class is specific to that class, linguistic analysis seems promising for future classification experiments.

Besides investigating automatic classification performance on the full, non-prototypical SIMIS data, future work will fore mostly need to add further labeller tracks to the resource and deal with suited ways to find a mapping to less complex tasks facing the 'full realism' of noisy real-life speech.

7. Acknowledgement

The authors would like to thank the student assistants Judith Köppe and Jürgen Glatz for their assistance during data collection.

8. References

M.E. Allaf, S.V. Jackman, P.G. Schulam, J.A. Cadeddu, B.R. Lee, R.G. Moore, and L.R. Kavoussi. 1998. Laparoscopic Visual Field. Voice vs Foot Pedal Interfaces for Control of the AESOP Robot. In *Surg. Endosc. 12 (12)*, pages 1415–1418.

F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber. 2005. An emotion-aware voice portal. In *Proc. Electronic Speech Signal Processing ESSP*.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

L. Devillers, L. Vidrascu, and L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks – Special Issue on "Emotion and Brain"*, 18(4):407–422.

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.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 488–500, Berlin-Heidelberg. Springer.

J.H.L. Hansen and S. Bou-Ghazale. 1997. Getting started with susas: A speech under simulated and actual stress database. In *Proc. EUROSPEECH-97*, volume 4, pages 1743–1746, Rhodes, Greece.

R. Hurteau, S. DeSantis, E. Begin, and M. Gagner. 1994. Laparoscopic surgery assisted by a robotic cameraman: concept and experimental results. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 2286–2289.

V.F. Munoz, C. Vara-Thorbeck, J.G. DeGabriel, J.F. Lozano, E. Sanchez-Badajoz, A. Garcia-Cerezo, R. Toscano, and A. Jimenez-Garrido. 2000. A medical robotic assistant for minimally invasive surgery. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 2901–2906.

B. Schuller, G. Rigoll, S. Can, and H. Feussner. 2008. Emotion sensitive speech control for human-robot interaction in minimal invasive surgery. In *Proc. 17th Intern. Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, pages 453–458, Munich, Germany. IEEE.

B. Schuller, S. Can, H. Feussner, M. Wöllmer, D. Arsic, and B. Hörnler. 2009a. Speech control in surgery: a field analysis and strategies. In *Proc. ICME*, pages 1214–1217, New York, NY, USA. IEEE.

B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu. 2009b. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. *Image and Vision Computing Journal , Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 27:1760–1774.

B. Schuller, S. Steidl, and A. Batliner. 2009c. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*, pages 312–315, Brighton, UK. ISCA.

S. Steidl. 2009. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin. PhD thesis.

S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. 2002. Development of user-state conventions for the multimodal corpus in SmartKom. In *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 33–37, Las Palmas.

D. Ververidis and C. Kotropoulos. 2003. A review of emotional speech databases. In *Proc. Panhellenic Conference on Informatics (PCI)*, pages 560–574, Thessaloniki, Greece.

Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.

A corpus for identification of speakers and their emotions

Marie Tahon, Agnès Delaborde, Claude Barras, Laurence Devillers
LIMSI-CNRS
BP 133, 91 403Orsay cedex, France
E-mail: (mtahon | agdelabo | barras | devillers)@limsi.fr

Abstract

This paper deals with a new corpus, called corpus IDV for "Institut De la Vision", collected within the framework of the project ROMEO (Cap Digital French national project founded by FUI6). The aim of the project is to construct a robot assistant for dependent person (blind, elderly person). Two of the robot functionalities are speaker identification and emotion detection. In order to train our detection system, we have collected a corpus with blind and half-blind person from 23 to 79 years old in situations close **to** the final application of the robot assistant. This paper explains how the corpus has been collected and shows first results on speaker identification.

1. Introduction

The aim of the project ROMEO[1] is to design a robotic companion (1m40) which can play different roles: a robot assistant for dependent person (blind, elderly person) and a game companion for children. The functionalities that we aim to develop are speaker identification (one speaker among N, impostor) and emotion detection in every day speech. The main challenge is to develop a strong speaker detection system with emotional speech and an emotion detection system knowing the speaker. All our systems are supposed to be real time systems.

In the final demonstration, the robot assistant will have to execute some tasks as defined in a detailed scenario. The robot is in an apartment with its owner, an elderly and blind person. During the whole day, the owner will have some visitors. The robot will have to recognize who are the different characters: his little children (two girls and a boy), the doctor, the house-keeper and an unknown person. In the scenario the robot will also have to recognize emotions. For example, Romeo would be able to detect how the owner feels when he wakes up (positive or negative) and to detect anger in the little girl's voice.

To improve our detection systems (speaker and emotion) we need different corpora, the closer to final demonstration they are, the better the results will be. We focused on blind or half-blind speakers (elderly and young person) and children voices while they interact with a robot (Delaborde et al., 2009) in order to have real-life conditions. However, emotions in real-life conditions are complex and the different factors involved in the emergence of an emotional manifestation are strongly linked together (Scherer, 2003).

In this paper, we will describe the IDV corpus which was collected with blind and half-blind person: acquisition protocol, scenarii involved. Then we explain the annotation protocol. And in section 4, we give our first results on speaker identification (identify a speaker from a set of known speakers).

2. IDV corpus

The part of the final scenario that concerns IDV corpus, we aim to demonstrate at the end of the project consists in:

- identify a speaker from a set of known speakers (children or adults),
- recognize a speaker as unknown and in this case, provide its category (children, adult, elderly) and gender (for adults only),
- and detect positive or negative emotion.

Speaker identification and emotion detection are real time tasks. For that objective, we have collected a first corpus called IDV corpus with blind and half-blind French people from 23 to 79 years old. This corpus has been collected without any robot but a Wizard-of-oZ which simulates an emotion detection system. This corpus is not fully recorded yet; further records are scheduled with the IDV. A second corpus will be collected in the context of the scenario: at the IDV (Institut de la Vision in Paris) with the robot ROMEO.

2.1 Corpus characteristics

So far, we recorded 10h48' of French emotional speech. 28 speakers (11 males and 17 females) were recorded with a lapel-microphone at 48kHz.

In accordance with the Romeo Project target room, the recordings took place in an almost empty studio (apart from some basic pieces of furniture), which implies a high reverberation time.

The originality of this corpus lies in the selection of speakers: for a same scientifically controlled recording protocol, we can compare both young voices (from 20 years old) to voices of older person (so far, the oldest in

[1] Cap Digital French national project founded by FUI6, http://www.projetromeo.com

this corpus is 89).

2.2 Acquisition protocol

Before the recording starts, the participant is asked some profile data (sex, location, age, type of visual deficiency, occupation and marital status). An experimenter from the LIMSI interviews the volunteer following three sequences described below in 2.3.

Some parasite noise happened to be audible in the studio (guide dog walking around, people working outside, talking, moving in the corridor, etc.). When overlapping the speaker's speech, these parts were discarded.

2.3 Sequences description

Each recording is divided into three sequences. The first one is an introduction to the Romeo project: we explain the participant that we need him to provide us with emotional data, so that we can improve our emotion detection system in a future robot. We take advantage of this sequence to calibrate the participant's microphone. Since there is no experimental control over the emotions that could be expressed by the participant, this part is discarded in the final corpus and will not be annotated.

In the second sequence, called “words repetition” (table 1), the experimenter asks the participant to repeat after him orders that could be given to the robot. The participant is free to choose the intonation and the expression of his or her production. This sequence gives us a sub-corpus where lexicon is determined and emotions mainly neutral.

Viens par ici! (Come over here!)	Mets le plat au four! (Put the dish into the oven!)
Arrête-toi! (Stop there!)	Descends la poubelle! (Put the bin out!)
Stop!	Va chercher le courrier! (Go and bring back the mail!)
Ecoute-moi! (Listen to me!)	Va chercher à boire! (Bring me something to drink!)
Approche! (Come closer!)	Aide-moi à me lever! (Help me to get up!)
Va-t-en! (Go away!)	Aide-moi à marcher! (help me to walk!)
Donne! (Give it!)	Ramasse ça! (Pick that up!)
Roméo, réveille-toi! (Romeo, wake up!)	

Table 1: List of words and expressions in French

In the third sequence, called “scenarii”, the experimenter presents six scenarii (see table Scenarii) in which the participant has to pretend to be interacting with a domestic robot called Romeo. For each presented scenario, the experimenter asks the participant to act a specific emotion linked to the context of the scenario : for instance Joy, "Your children come to see you and you appreciate that, tell the robot that everything is fine for you and you don't need its help", or Stress, "You stand up from your armchair and hit your head in the window, ask Romeo to come for help", or Sadness, "You wake up and the robot comes to ask about your health. You explain it that you're depressed". The participant has to picture himself or herself in this context and to speak in a way that the emotions are easily recognizable. He (she) knows that the lexicon he (she) uses is not taken into account; the emotion has to be heard in his or her voice.

At the end of each of his or her performance, the experimenter runs a Wizard-of-Oz emotion detection tool that tells aloud the recognized emotion. The system is presented as being under-development, and most of the times it does not correctly recognize the emotion: it can recognize an emotion that is of the opposite valence of what the participant was supposed to express (the experimenter selects Anger when Joy has been acted); it can recognize no emotion at all (the experimenter selects Neutral when a strong Anger was expressed, or when the emotion has not been acted intensely enough); it can recognize an emotion that is close to what is expected, but too strong or too weak (Sadness instead of Disappointment). The participant is asked to act the emotion again, either until it is correctly recognized by the system, or when the experimenter feels that the participant is tired of the game.

Emotional data acquired through acting games obviously do not reflect real-life emotional expressions. However, the strategies that are being used through our Wizard-of-Oz emotion detection tool allow us to elicit emotional reaction in the participants. An example: the participant is convinced that he expressed Joy, but the system recognizes Sadness. The participant's emotional reactions are amusement, frustration, boredom or irritation.

Our corpus is then made of both acted emotions, and spontaneous reactions to controlled triggers. The distinction between acted and spontaneous expressions will be spotted in our annotations; this distinction is really important to have an estimation of how natural the corpus is (Tahon & Devillers, 2010).

We can also question the relevancy of having the participant imagine the situation, instead of having him live it in an experimental setting. We should note that for obvious ethical reasons we cannot put them in a situation of emergency such as "being hurt, and ask for immediate help": we can only have them pretend it.

Another obvious reason for setting this kind of limited protocol is a matter of credibility of the settings: currently, the only available prototype does not fit the target application characteristics (Nao is fifty centimeters high, and its motion is still under development).

Scenarii	Emotions
Medical emergency	Pain, stress
Suspicious noises	Fear, anxiety
Awaking (good mood)	Satisfaction, joy
Awaking (bad health)	Pain, irritation, anger
Awaking (bad mood)	Sadness, irritation
Visit from close relations	Joy

Table 2: Scenarii

Table 2 summarizes the 6 different scenarii and the emotions asked to the participant.

3. Corpus annotations

3.1 Affective state labels

Segmentation and annotation of the data are done with the Transcriber annotation tool[2] on the scenario sequences.

The participant utterances are split into emotional segments. These segments mark the boundary of the emotion: when a specific emotion expression starts, and when it comes to an end.

On each segment, three affective state labels describe the emotion. Affective state labels include emotions, attitudes and communication acts. The first label corresponds to the most salient perceived affective state, while the two others characterize more precisely the emotion, balance it. The table Affective state labels presents the emotional annotation values that are used. The category "others" presents labels that are not classified in macro-classes.

Other dimensions are annotated:

- Intensity: the strength of the emotion, 5 scales from *very weak* to *very strong*.
- Activation: how many different phonatory means are involved to express the emotion (voice trembling, change in loudness…), 5 scales from *very few* to *a lot*.
- Control: does the speaker contain the expression of the emotion, 5 scales from *not at all* to *completely*.
- Valence: does the speaker feel positive or negative ? *positive*, *negative*, *positive and negative*, *either positive or negative*, *valence* indeterminable.
- Audio quality: if the recorded segment quality is fine or not (microphone noise, participant speaking too close…), from *good* to *bad*.
- Spontaneous/Acted : a simple flag meant to spot if the participant was at that time acting an emotion in the context of a scenario, or reacting spontaneously to an event.

[2] http://trans.sourceforge.net/en/presentation.php

Affective state categories	Annotations values
POSITIVE	Joy
	Amusement
	Satisfaction
	Positive
	Motherese
ANGER	Anger
	Irritation
SADNESS	Sadness
	Disappointment
FEAR	Fear
	Anxiety
	Stress
	Embarrassment
NEUTRAL	Neutral
OTHERS	Irony
	Compassion
	Interest
	Scorn
	Boredom
	Empathy
	Pain
	Excitation
	Surprise
	Negative
	Overbid
	Provocation

Table 3: Affective state labels. Others are not classified in macro-classes, and include affective and mental states, and communication acts.

3.2 IDV emotional content

Two expert labelers perform the annotation. As the emotional annotation of the IDV corpus is not finished yet, all results on emotion annotation are based on a set of 15 speakers.

IDV corpus is divided into two different corpora: spontaneous and acted, according to the task (as defined in part 3). The results of the emotion scores are reported in table 4.

The spontaneous corpus contains 736 instances of 0.5s to 5s. The most important emotional label is “interest” (50.5%). This corresponds to the agreement of the volunteer with what the interviewer asked him to do. Positive emotions (18.4%) are more numerous than Anger, Sadness and Negative emotions (7.8%). The volunteers have accepted to be recorded, so they were not supposed to express displeasure, they will more probably be nice with the LIMSI team.

Macro-class “fear” (mainly anxiety) is also quite important (9.5%). It corresponds to embarrassment or anxiety, playing the actor is not an easy task. Macro-class “boredom” is only 2.1%.

The acted corpus contains 866 instances of 0.5s to 6s. The results correspond to what was expected: the main

emotions are well represented. Positive emotion (21.2%, mainly "satisfaction"), Anger emotion (18.2%, mainly "irritation"), Fear (24.2%, mainly anxiety) and Sadness (8.3%, "disappointment" and "sadness").

Label emotion	Spontaneous	Acted
Joy	0.82	5.54
Amusement	8.49	0.58
Satisfaction	4.89	10.51
Positive	3.94	4.16
Motherese	0.27	0.46
POSITIVE	**18.41**	**21.25**
Anger	0.2	3.23
Irritation	1.9	15.01
ANGER	**2.1**	**18.24**
Sadness	1.63	3.98
Disappointment	2.72	4.33
SADNESS	**4.35**	**8.31**
Fear	0.07	4.97
Anxiety	6.66	16.28
Stress	1.56	2.89
Embarrassment	1.22	0.06
FEAR	**9.51**	**24.2**
Neutral	7.2	2.25
NEUTRAL	**7.2**	**2.25**
Irony	0.07	0.23
Boredom	2.17	1.04
Negative	1.36	4.73
Surprise	4.14	2.37
Pain	0.07	3.87
Excitation	0.07	0.46
Interest	50.54	12.93
OTHERS	**58.42**	**25.63**

Table 4: Emotion scores (%) for both spontaneous and acted IDV corpora

4. IDV first results

In this section, speaker identification scores are presented. All the results presented here were obtained with the same method based on GMM (Gaussian Mixture Models) speaker models (Reynolds et al., 2000). First we have studied the different parameters of the GMM model, then the evolution of scores in function of the sex and the age of speakers.

4.1 Global speaker identification scores

This section aims at choosing the experimental setup for studying the influence of the age, gender and emotional expression. Experiments are performed with the "repeating words" sequence of the corpus. It contains 458 audio segments of varied duration. 26-dimensional acoustic features (13 MFCC and their first-order temporal derivatives) are extracted from the signal every 10ms using a 30ms analysis window.

For each speaker, a training set is constructed by the concatenation of segments up to a requested duration Ntrain; a Gaussian mixture model (GMM) with diagonal covariance matrices is then trained on this data through maximum likelihood estimation with 5 EM iterations. The remaining segments, truncated to a Ntest duration, are used for the tests. For a given duration, the number of available segments is limited by the number of segments already used for training and the minimal test duration necessary (the higher duration is, the less audio files there are). For each test segment, the most likely speaker is selected according to the likelihood of the speaker models.

In order to optimize the number of files of train and test, we have chosen the following set of parameters:

- test duration: 1s (225 files),
- train duration: 10s (179 files),
- speaker model: mixture of 6 Gaussians.

The error rate is 34.7% (+/- 6.5%) when recognizing one speaker among 28.

This extremely short test segment duration is due to constraints on segment counts in the database, and improvement of the performance as a function of the segment length will be studied later in the course of the project.

4.2 Age influence

In this part, we show that speaker identification is easier on elderly person voices than on young voices. Two sub-corpora from IDV corpus composed of the 8 older volunteers (4 male, 4 female, from 52 to 79 years old), respectively the 8 younger volunteers (4 male, 4 female, from 23 to 46 years old) are studied separately. Of course, the number of segments is quite low, which may be a bias of the experiment.

The results are referred in the table 5, error rate, number of segments for test and trust interval (binomial distribution test).

	Old person	Young person
Error rate	17.00%	38.00%
Number of segments	66	63
Trust interval	9.18%	12.24%

Table 5: Speaker identification, age influence: error rate, number of segments and trust interval

As a result speaker identification (one speaker among N) is better with elderly person voices. Our hypothesis is that voice qualities are much more different with elderly person voices than with young voices. In figure 1, we have plotted the MFCC2 Gaussian model for the first four older person (blue) and for the first four younger person (red). As the red curves are quite the same, the blue one are more separated one from another.

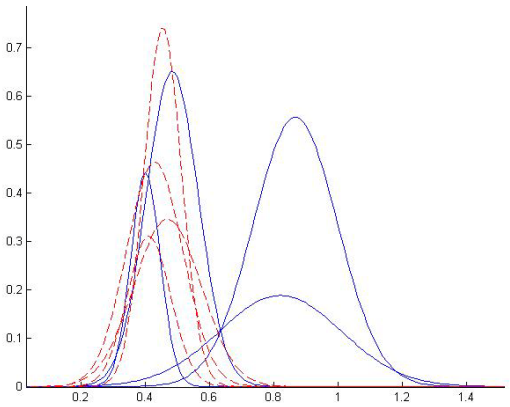

Figure 1: Distribution of the 4th MFCC coefficient according to a Gaussian model for old (plain) and young speaker (dashed)

4.3 Sex influence

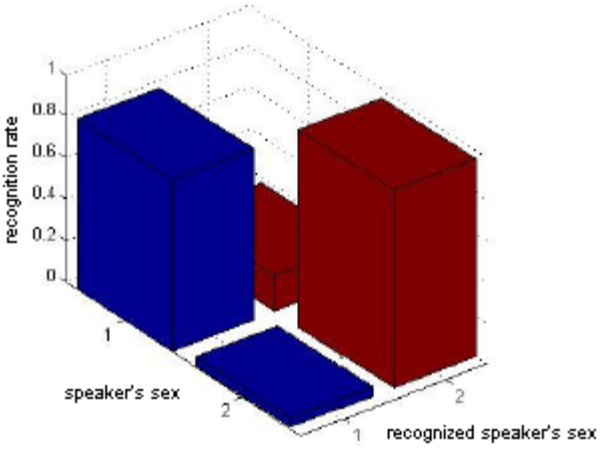

Figure 2 : Confusion matrix between male (1) and female (2)

Based on the whole IDV corpus, we compute the confusion matrix sorted by sex without taking into account the age of the speakers anymore.

A female voice is recognized as well at 96% (in Fig.2, the female-female block), a male voice is recognized as well at 82% (male-male block). Female voices have better identification scores.

4.4 Emotional speech influence

The results below are based on the corpus "repeating words" which contains 28 speakers. The results presented in this part are based on both sequences "repeating words" and "scenario", with the 15 speakers corresponding to the emotional annotation of the sequence "scenario". Table 5 below shows the error rate for speaker identification (1 among 15) across the 3 corpora: "repeating words", "scenario spontaneous" and "scenario acted".

The parameters we have chosen for the Gaussian model are the followings: 5 Gaussians, train duration: 10s, test duration: 1s.

		TEST		
		"Words"	**"Spontaneous"**	**"Acted"**
TRAIN	**"Words"**	28.60%	78.60%	88.00%
	"Spontaneous"	X	45.10%	60.20%
	"Acted"	X	X	56.30%

Table 5: Error rates for speaker identification across the three corpora

Identification scores are better with the "words" corpus (lexically controlled) than with the "acted" corpus. The "spontaneous" corpus gives intermediate results. The scores are always better when the train and the test are made on the same corpus.

Speaker models were tested directly in mismatch conditions without any specific adaptation. The very high error rates observed are of course due to the very short train and test durations constraints in our experiments, but also highlight the necessity of an adaptation of the speaker models to the emotional context which will be explored during the ROMEO project.

5. Conclusion

This corpus IDV is interesting for many reasons. First, as it presents a sequence of words, lexically determined by the protocol and quite neutral, and a sequence of emotional speech, with the same speakers, recorded in the same audio conditions, it allows us to compare scores for speaker identification between neutral speech and emotional speech.

Secondly, the corpus collection has been made with blind and half-blind volunteers from 23 to 79 years old. Thus we can compare scores across speaker age. Moreover we have the opportunity to work with elderly person who often have specific voice qualities.

6. References

Delaborde A., Tahon M., Barras C., Devillers L. (2009). A Wizard-of-Oz game for collecting emotional audio data in a children-robot interaction. AFINE 2009.

Tahon M. and Devillers L. (2010). Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Proceedings of the Fifth International Conference on Speech Prosody*, 2010.

Devillers L., Abrilian S., Martin J-C., (2005) Representing Real-life Emotions in Audiovisual Data with Non Basic Emotional Patterns and Context Features, ACII 2005.

Devillers, L. Vidrascu, L. Lamel, L. (2005). Emotion detection in real-life spoken dialogs recorded in call center, *Neural Networks*, Special Issue on "Emotion and Brain", ELSEVIER, Vol. 18, No. 4, pp. 407-422, 2005

Reynolds D., Quatieri T., and Dunn R. (2000) Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

Scherer K. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication*, vol. 40, pp227-256, 2003;

Interfacing WORDNET-AFFECT with OCC Model of Emotions

Alessandro Valitutti and Carlo Strapparava

FBK-irst, 38050, Povo, Trento, Italy

Abstract

In this paper we presented WNAFFECT-OCC, an extension of WORDNET-AFFECT lexical database (Strapparava and Valitutti, 2004) consisting of the integration between the hierarchy of affective labels and a specific appraisal model of emotions (i.e. OCC - Ortony Clore and Collins - model (Ortony et al., 1988)) widely employed in computational applications. In OCC, emotions are classified according to some categories, typically employed in the appraisal process, such as events, objects, and actions. WNAFFECT-OCC can be exploited to perform improvements in sentiment analysis and affect sensing.

1. Introduction

Affective lexical resources are largely employed in sentiment analysis and affect sensing of text. The reason is that part of the affective meaning is represented at the lexical level. To recognize the emotional state expressed in a text, it is necessary to detect the relation between words and emotions. In spite of that, in several cases there are ambiguities that cannot be resolved without an analysis of the semantic context expressed at the sentence level.

According to (Strapparava et al., 2006), there are two types of affective words: *direct* and *indirect affective words*. The former type includes words that directly refer to some specific emotional state (e.g. 'fear' or 'enjoyed'). The latter type includes all the other words.

While direct affective words are not ambiguous (according to their definition), Indirect affective words can be characterized by affective lexical ambiguity, and then can refer to different possible emotions. For example, words such as 'help', 'revenge', 'victory', or 'lottery' are positive or negative (i.e., associated to emotions with positive or negative valence) according to the *role of the subjects* taken in account in the textual analysis. For example, the word 'victory' can be used to express *pride*, if communicated by the 'winner'. On the other hand, if the author of the message is the 'loser', the expressed emotion is *frustration* or *disappointment*. Then a word can refer to different emotions according to the subject semantically connected to it.

Another type of contextual information is of the temporal type. Depending on whether an event is localized in the past or in the future, the emotion evoked by it can be different. If the 'victory' is a future (and desired) event, the expressed emotion is hope. On the other hand, if the event is already happened, a possible emotion is *satisfaction*.

In order to overcome the ambiguity of the indirect affective words, we have take in account the appraisal theories of emotions, according to which emotions are induced by a process of cognitive evaluation of perceived conditions.

Then we performed an extension of WORDNET-AFFECT lexical database (Strapparava and Valitutti, 2004) consisting of the integration between the hierarchy of affective labels and a specific appraisal model of emotions (i.e. OCC - Ortony Clore and Collins - model (Ortony et al., 1988)) widely employed in computational applications. In OCC, emotions are classified according to some categories typically employed in the appraisal process. The main categories are: events, objects, and actions.

The information from this model can be detected from the textual context of the word to disambiguate, and can refer to different subject or roles in an action, or different temporal conditions.

The development of the resource presented in this work (called WNAFFECT-OCC) consisted of two stages:

1. OCC emotional taxonomy was rearranged in order to emphasize the relation between emotions with the same type but opposite polarity.
2. OCC was interfaced with the hierarchy of WorNet-Affect.

2. WORDNET-AFFECT

WORDNET-AFFECT is an extension of WORDNET database (Fellbaum, 1998), including a subset of synsets suitable to represent affective concepts. Similarly to the "domain label" methodology (Magnini and Cavaglià, 2000), one or more affective labels (*a-labels*) were assigned to a number of WORDNET synsets. In particular, the affective concepts representing emotional states (and including direct affective words) are identified by synsets marked with the a-label EMOTION. There are also other a-labels for those concepts representing moods, situations eliciting emotions, or emotional responses. WORDNET-AFFECT is freely available for research purpose at `http://wndomains.itc.it`. See (Strapparava and Valitutti, 2004) for a complete description of the resource.

WORDNET-AFFECT is supplied with a set of additional a-labels (i.e. the emotional categories), hierarchically organized, in order to specialize synsets with a-label EMOTION. Emotional categories are classified according to vaules of emotional valence, represented by four additional a-labels: POSITIVE, NEGATIVE, AMBIGUOUS, and NEUTRAL. The first one corresponds to "positive emotions", defined as emotional states characterized by the presence of positive edonic signals (or pleasure). It includes synsets such as `joy#1` or `enthusiasm#1`. Similarly the NEGATIVE a-label identifies "negative emotions" characterized by negative edonic signals (or pain), for example `anger#1` or `sadness#1`. Synsets representing affective states whose

valence depends on semantic context (e.g. `surprise#1`) were marked with the tag AMBIGUOUS. Finally, synsets referring to mental states that are generally considered affective but are not characterized by valence (e.g. `apathy#1`), were marked with the tag NEUTRAL.

3. Integration of OCC Model

OCC model of emotions is based on a theory of emotions according to which emotional states arise from the cognitive appraisal of a perceived situation. OCC can be viewed as an ontology of concepts (i.e. objects, events, actions) expressing possible causes of emotions, and represented in text as object of the discourse. We rearranged OCC and integrate it with the affective hierarchy of WORDNET-AFFECT. In this way we obtained a new hierarchy with a set of additional labels (called *OCC labels*). Part of them represent OCC emotions and are interfaced with the existing emotional categories of WORDNET-AFFECT. The remaining OCC labels denotes emotional types representing conditions eliciting emotions. For example, the label HAPPY-FOR represents an OCC emotion and is interfaced with the a-label JOY. The parent label in the hierarchy is FORTUNE and characterizes emotions generated by the evaluation of positive events happened some other people.

First OCC hierarchy of emotional types were re-arranged in order to clearly distinguish between positive and negative emotions with the same type of appraisal conditions. In a second stage, each OCC emotion were associated (as additional attributes) to a corresponding category of WordNet-Affect, and consequently to all its subcategories in the hierarchy. In this way, we connected the 24 OCC emotions with 179 of the 382 categories of the affective hierarchy. The remaining 203 categories are subnodes of the other and then inherit the OCC information. Emotions marked in the hierarchy as ambiguous (i.e. "surprise") or neutral (i.e. "apathy") were not interfaced (because there are not corresponding categories in OCC).

Table 1 shows the emotions elicited by the appraisal of positive (*fortune*) or negative (*misfortune*) conditions of other people. Each condition is associated to two possible emotions, according to opposite values of valence. Thus polarity of emotions is associated, with all possible combinations, to polarity of conditions. Table 2 shows the association between valence and appraisal factors in the case of emotions emerging from the evaluation of actions. In this case, it is important to identify the subject of the emotion because the same action can be associated to different emotions according to corresponding subjects. For example, a positive action generates *pride* in the actor, *gratitude* in the actee, and *admiration* in an external observer. In Figure 1, Figure 2, and Figure 3 the main subtrees of the OCC model, interfaced with WORDNET-AFFECT, are shown.

	Positive Emotions	*Negative Emotions*
fortune	happy-for	envy
misfortune	gloat	pity

Table 1: Emotions arising from the evaluation of "fortune/misfortune" of other people.

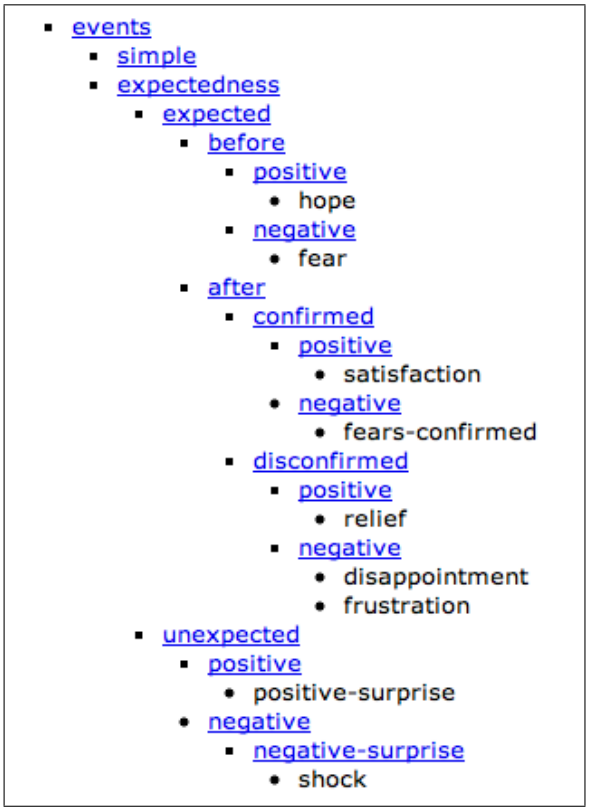

Figure 1: Emotions arising from the appraisal of "expectedness of events".

Figure 2: Emotions arising from the appraisal of "actions".

	Positive Emotions	*Negative Emotions*
actor	pride	shame
actee	gratitude	resentment
observer	admiration	reproach

Table 2: Emotions arising from the evaluation of actions.

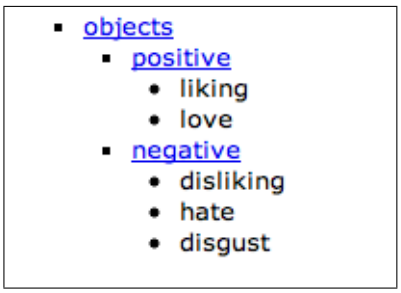

Figure 3: Emotions arising from the appraisal of "objects".

In the OCC model, there are concepts that can be related to different emotions, according to different conditions inducing emotion. For instance, the action help can generate satisfaction in the subject of the action (actor), gratitude in the object of the action (actee) and admiration in the people that evaluate the action from an external point of view (observers).
After rearranged the affective hierarchy, we tagged the 526 indirect synsets of WORDNET-AFFECT, with Part of Speech *noun*. Each synset was associated to an *ontological category* (i.e. event, object, action, attitude, trait) selected from its hypernyms, to one or more *subjects* (e.g., in the case of actions, corresponding to the possible point of view of *actor*, *actee* and *observer*), and one o more emotions selected from the affective hierarchy. Although the annotation was performed at the synset level, we observed more general regularities that we aim to take in account for a further refining. For example, synsets denoting negative actions corresponds to positive/negative emotions in the actor (e.g. *pride/shame*), negative emotions in the actee (e.g. *distress*), and positive/negative emotions in the observer (e.g. *admiration/reproach*).

4. Synset Tagging

The enrichment of the affective hierarchy through the introduction of the OCC model allowed us to perform an annotation of a number of synsets containing indirect affective words. The schema of tagging is different from that employed for the direct affective words, and can be characterized as follows:

1. Each synset can be tagged with more emotion labels.
2. A number of additional labels, expressing appraisal attribute, are associated to the emotion labels.

For example, the synset {*contempt#n, scorn#n*} refers to an action; it is associated to the attributes "actor", "actee", and "others" (according to the subject experiencing that action). In turn, according to OCC model, the related emotions are respectively *disliking* ("actor"), *distress* and *disappointment* ("actee"), and *disapproval* ("others"). In this way, the individuation of the appraisal condition identified in the sentiment analysis of a text allows us to restrict the set of possible emotions associated to a specific concept.

```
<noun-syn id="n#05016741"
          synonyms="encouragement"
          category="behaviour">
  <actor emotion="compassion"/>
  <actor emotion="admiration"/>
  <actee emotion="gratitude"/>
  <actee emotion="pride"/>
  <actee emotion="satisfaction"/>
  <actee emotion="positive-hope"/>
</noun-syn>

<noun-syn id="n#04005196"
          synonyms="vanity emptiness"
          category="quality">
  <subj emotion="pride"/>
  <others emotion="reproach"/>
  <others emotion="dislike"/>
</noun-syn>

<noun-syn id="n#00270022"
          synonyms="cruelty inhuman_treatment"
          category="action">
  <actor emotion="satisfaction"/>
  <actor emotion="anger"/>
</noun-syn>

<noun-syn id="n#03790984"
          synonyms="ambition ambitiousness"
          category="trait">
  <subj emotion="pride"/>
  <subj emotion="enthusiasm"/>
  <others emotion="admiration"/>
  <others emotion="envy"/>
</noun-syn>
```

Figure 4: Some instances of synsets tagged according to OCC model. Synonyms (associated in WORDNET to each synset ID) are shown in this example.

Given the complexity of the semantic connection between indirect affective words and emotions, the annotation of synsets is necessarily a work in progress. We tagged a first set of 526 synsets with PoS *noun*. At the moment we focus on a specific appraisal attribute (called *subjectivity*), denoting the subject experiencing the expressed emotion. As in the above example, the possible values are "actor", "actee", and "others". We consider not only synsets referring to actions, but also attitudes, events, and every elements perceived as possible cause of emotions.
The reason of the choice of subjectivity is related to the state-of-the-art in sentiment analysis, in which the identification of the subject (or *emotion/opinion holder*) is a crucial aspect of the affective sensing. More generally, the introduction of new tags has to be performed according to the current capability to extract the necessary contextual information. For example, the effectiveness in the recognition of "expectedness" of events is related to the capability to perform a temporal analysis. In Figure 4 shows some synsets with the corresponding annotation.

5. Possible Applications

WNAFFECT-OCC can be exploited to perform improvements in sentiment analysis and affect sensing. Specifically, two different but mutually related tasks can take advantage of this resource. The one is recognition of the affective state expressed in the text. The other one is the polarity value (*positive* vs. *negative*) of emotions, opinions, events, or objects, according to different criteria. We call *emotion-holder* the subject experiencing the emotion and *emotion-target* the object of the emotion.

A possible procedure for the automatic recognition of the expressed emotion consists of the following steps:

1. identification of the emotion-holder
2. identification of the emotion-target
3. OCC analysis of the emotion-target

The first two steps are a standard in sentiment analysis. We make the assumption that the emotion-target is also the source of the emotion through the appraisal process. The reason is that it is a common experience to identify, in the description of emotional experience, the cause and the object of the emotion. This assumption allows us to apply OCC model (integrated in the affective hierarchy) to above procedure.

A key characteristic of the proposed use of WNAFFECT-OCC is the connection of the polarity of indirect words and the polarity of emotional valence. For example, in the sentence "I believe that you'll got that job", the word 'belief' is affectively ambiguous, but can be interpreted as expressing *hope* because the expected event is typically positive. On the other hand, in the sentence "I am worried that George will come to the party" the expressed emotion is *worry*; therefore the expected event is negative.

Another example is related to the word 'help' denoting an action that can be positive or negative according to different roles and temporal conditions. In "I need help", the author is the "actee" and time is "before" the action. Then the expressed emotion is recognized as *sadness*. Instead in "Your help was useful", the time is "after" the action, and thus the expressed emotion is *gratitude*.

Finally, WNAFFECT-OCC can be used to identify the expressed emotion in texts containing both positive and negative words. For example, in "He is too damn lucky!" there is a positive ('lucky') and a negative word ('damn'). The exploration of WNAffect-OCC hierarchy allows us to identify the following information:

1. The emotion-target is an event.
2. The event is a condition of others.
3. The event is positive, and then the emotion is elicited by "fortune of others".
4. The emotion is negative (by the evaluation "too damn"), and then the emotion is recognized as *envy*.

6. Conclusion

In this paper we presented WNAFFECT-OCC, an extension of WORDNET-AFFECT lexical database (Strapparava and Valitutti, 2004) consisting of the integration between the hierarchy of affective labels and a specific appraisal model of emotions (i.e. OCC - Ortony Clore and Collins - model (Ortony et al., 1988)) widely employed in computational applications. In OCC, emotions are classified according to some categories, typically employed in the appraisal process, such as events, objects, and actions. WNAFFECT-OCC can be exploited to perform improvements in sentiment analysis and affect sensing.

In this work we employed the distinction between direct and indirect type of affective lexicon. The former is characterized by a simple denotative connection with emotional categories. In the latter, the association can be more complex. We selected a specific context and focus on words denoting actions, events, and all types of stimuli inducing emotions through the appraisal process. This choice allows us to employ a specific model of appraisal as a way for improving the organization of affective lexicon. Finally we emphasize the intrinsic ambiguity of this type of words and the role of contextual information. We believe that the integration of lexical information represented in WNAFFECT-OCC and the contextual information (detected by different strategies) is a promising way for improving the performance of sentiment analysis techniques.

7. References

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

B. Magnini and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *Proc. of the 2^{nd} International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece.

A. Ortony, G. L. Clore, and A. Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, New York.

C. Strapparava and A. Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proc. of 4^{th} International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, May.

C. Strapparava, A. Valitutti, and O. Stock. 2006. The affective weight of lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, May.

Linguistic Structure, Narrative Structure and Emotional Intensity

Tibor Pólya[1], Kata Gábor[2]
Institute for Psychology, Hungarian Academy of Sciences[1]
Linguistics Institute, Hungarian Academy of Sciences[2]
polya@mtapi.hu, gkata@nytud.hu

Abstract

This paper presents a Hungarian corpus of emotional speech, constructed to promote the study of the supposed correlation between emotional intensity and narrative structure. According to our hypothesis, narrators construct a more elaborated narrative structure when the intensity of their emotional experience is higher. This claim is based on the narrative psychological approach, according to which subjective experiences are organized into a narrative structure. In order to verify and refine this hypothesis, a corpus of recorded Hungarian speech data was created and annotated on three distinct levels: standard psycho-physiological measures of emotional intensity, relevant features of narrative structure, and linguistic markers associated to narrative structural features. In parallel to this research, we aim to develop a tool to automatically analyze and annotate linguistic markers of narrative structure. This tool is implemented as a part of the Hungarian module for the corpus processing environment NooJ.

1. Introduction

The present paper describes the construction of a Hungarian corpus of emotional speech, annotated on three distinct levels related to emotional intensity: standard psycho-physiological measures, narrative structure, and linguistic realization of narrative structural categories. The project aims to explore correlations between the emotional state of the narrator and the narrative structure of verbal reports on emotional episodes. Furthermore, we construct algorithms for the automatic identification and categorization of the linguistic markers related to the narrative structure. The annotated corpus serves as a basis for the elaboration of an algorithm to automatically identify emotional intensity in written language corpora, based on a mapping between local syntactic structures and the narrative structure of the text.

A frequently used method for studying emotional experience is to ask people about emotional episodes. According to Stein and her colleagues (e.g. Stein & Hernandez, 2007) the content of these verbal reports is informative on emotional processes, since verbal reports indicate how people understand their emotional experiences. On the other hand, verbal reports on emotional episodes frequently have a narrative structure. We argue that narrative organization is an important feature of these verbal reports and that narrative structure correlates with the intensity of emotional experience. Our hypothesis is that narrators build up a more elaborated narrative structure when the intensity of their emotional experience is higher compared to cases when their emotional experience is less intensive.
This argument can be based on two lines of research. First, there is considerable empirical research showing that narrative structure has a significant effect on readers' and beholders' emotions (Brewer & Lichteinstein, 1982; Oatley, 1999; Tan, 1996). Second, the narrative psychological approach claims that subjective experiences are organized into a narrative structure (Bruner, 1990; 2008; Sarbin, 1986). László (2008) argues that the psychological relevance of narrative structural categories can be discerned. Based on the above lines of research, we expect to be able to infer the intensity of the narrator's current emotional state from the structure of an emotional self-narrative.

Several computational tools have been developed to automate or facilitate the analysis of self-narratives. However, the capacities of these pieces of software are limited to the analysis of the content. The majority of these tools only perform a lexical analysis, i.e. counting word occurrences and comparing frequencies of words belonging to different lexical categories (e.g. Bucheim, & Mergenthaler, 2000). Our attempt is significantly different in that we make use of a much deeper linguistic analysis (constituent parsing), and beside lexical categories our model builds on structural properties of the texts.

2. The corpus

2.1 Recording and preprocessing

The corpus is composed of recorded Hungarian speech data from the participants and the written transcription of the texts. There were 60 healthy adult people participating in the project, their age varied between 18 and 45 years (M=29.2 SD=6.3).
The following psycho-physiological channels were registered by a PROCOMP5 biofeedback system: Blood Volume Pulse, Abdominal Respiration and Skin Conductance. First, the baseline of psycho-physiological measures was registered during a 3 minutes period. Then a cue word paradigm was used to elicit autobiographical memories regarding emotional episodes. Every participant has related four memories with the following cue words: proud, relief, sadness and fear. The three channels were continuously measured during narration.
The transcribed texts were then preprocessed using the corpus development environment NooJ (Silberztein, 2008). The preprocessing includes tokenization and

segmentation into narrative clauses.
To assess emotional intensity during recounting the following 5 psycho-physiological measures were used: Heart Rate Relative to Baseline, Heart Rate Amplitude Relative to Baseline, Abdominal Respiration Rate, Abdominal Respiration Amplitude and Skin Conductance Relative to Baseline. Measures were aggregated by narrative clauses.

2.2 Annotation of narrative structure

A pre-defined set of characteristics of narrative structure were manually annotated on the level of narrative clauses. Each narrative clause belonging to the categories below was assigned one or more labels. The encoding of these properties is based on the coding scheme developed by (Pólya, Kovács, Gábor, & Kabai, 2008). This scheme provides a more detailed sub-division of the following categories:

1) *Embedded evaluation:* During narration, the speaker continuously evaluates the related events. Emotional reactions can be seen as a consequence of the person's evaluation on the event in question (Scherer, 2001).
2) *Temporal unfolding:* Emotional reaction is a transient mental phenomenon, and as such, it has a temporal unfolding (beginning, development, end point) (Stern, 1995).
3) *Subjectivity*: We can distinguish two formal variations of emotional experience: involvement and reflection (Lambie, & Marcel, 2002). At the level of narrative structure, this duality corresponds to the distinction between reliving and retrospective perspective forms.

3. Linguistic Analysis

3.1 Representation Levels

Narrative structure can be anchored to linguistic structural units. However, the above mentioned structural categories cannot be directly linked to linguistic entities: they constitute different levels of representation.

Besides studying the correlation between emotional intensity and narrative structure, we set the goal to produce a fine-grained linguistic description which would bridge the gap between textual data and the abstract representation level of narrative structure. The description has to be automatable in order to be integrated in an NLP tool chain which recognizes relevant narrative structural units in written language corpora.

3.2 Preprocessing and Syntactic Analysis

We opted for a rule-based system to analyze our corpus and develop a tool which automatically annotates narrative structural units. The first consideration behind this choice is that the size of the corpus does not allow building a statistical model suitable for our purposes (i.e. automatic annotation). Second, we believe that the linguistic embodiment of narrative structure is displayed at the level of morphosyntactic structure. Thus, a simple lexical analysis can only serve as an indicator in cases where a linguistic structure triggers the occurrence of some lexical units (typically function words), but direct correspondence can only be observed between complex syntactic structures and narrative categories. For instance, embedded evaluation is often expressed by a comparison: in such cases, we have to find not only the comparative adjective (or the pair of adjectives, in case of an opposition), but the scope of the comparison, which can be expressed by an NP or an embedded clause.

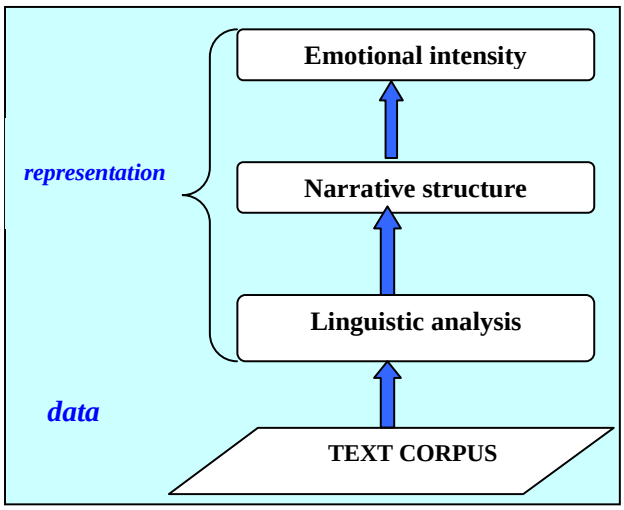

Fig.1.: Annotation levels in the corpus

For the algorithmic tasks we used the corpus processing and grammar editor environment NooJ (Silberztein, 2004). Besides being able to process and annotate corpora, NooJ allows to efficiently combine lexical, morphological and syntactic features. Moreover, its user friendly interface contributes to the sharing and re-use of grammars.
NooJ is an FST-based tool completed with additional functions (lexical constraints, feature unification, variables – see (Silberztein, 2008) for recent developments) which give it the descriptive power of a Turing machine. The finite state technology, which constitutes the computational background of NooJ, is appreciated for its efficiency: huge amounts of lexical data can be stored in a very compact form, and since input text can also be represented as a transducer, it makes grammar application an easy and fast operation. On the other hand, local grammars also comply with the linguistic approach according to which phrase structure rules that operate on grammatical categories often prove to be inadequate when confronted to "real" corpus data. The reason for this is that lexical items belonging to the same category frequently show different grammatical behavior, i.e. different distribution. Moreover, syntactic rules are far from being general among members of categories: they more frequently operate either on specific lexical items or on subclasses of parts of speech. As we will see in the following sections, this approach is increasingly applicable for narrative structure, where a mix of linguistic rules, heuristics and lexical triggers have

to be taken into consideration to provide the better coverage for spoken language data. What makes NooJ an especially favorable choice for developing such applications is its capacity to implement cascaded grammars in a way that the output of each processing step, as well as the original text itself, remains accessible during the whole processing chain. Its robustness and XML-compliance allows developing a complete tool chain for analyzing Hungarian corpora.

Hungarian linguistic resources for NooJ (dictionaries, syntactic grammars) are being developed at the Linguistics Institute of the HAS (Váradi, Gábor, 2004). The Hungarian module currently includes tokenization, sentence splitting, lemmatization, morphological analysis, clause boundary detection, constituent chunking and annotation of some basic dependency relations (e.g. auxiliaries and main verbs, detached verb prefixes etc.) To achieve better coverage, the internal lookup-based morphological analysis in NooJ was completed with the Humor morphological analyzer (Prószéky, 1995). The rule-based syntactic parser included in the Hungarian NooJ module is presented in details in (Gábor, 2007). The syntactic module is composed of a phrase chunker and a dependency annotator. The NP chunker developed by Váradi (2003) was integrated into it at an early stage. Chunking is performed by local grammars implemented as FSTs, with a high precision. Grammars range from very general syntactic (phrase-building) rules to very specific lexical patterns, exhibiting different degrees of generalization. Indeed, one of the most attractive features of local grammars in parsing is their flexibility: they allow for modeling lexically constrained collocations, semi-frozen expressions and syntactic rules with the same description method, i.e. with graphs representing NooJ transducers. The output of chunking is a labeled bracketing of the input sentence.
The next step of the syntactic analysis is clause boundary detection. Although narrative clauses were manually annotated in our corpus, the tool chain has to be able to recognize these structural units in previously unseen texts. Our clause boundary grammar consists of a set of FSTs representing grammar rules and heuristics, most of which are based on local phenomena such as conjunctions, punctuation marks and sentence adverbs.

The efficient modeling of some phenomena (e.g. postpositional phrases or auxiliaries) on the other hand benefits from NooJ's enhancements, especially the *lexical constraints*. In order to be able to make use of this function, Hungarian NooJ dictionaries were completed by a set of lexical syntactic features. They define finer grained distributional categories than parts of speech, making it possible to achieve higher precision than simple local grammars. Relevant dependency relations are those between verbs and auxiliaries, verbs and detached preverbs, postpositions and noun phrases. Auxiliaries and preverbs are particularly relevant since they show a strong correlation with *temporal unfolding*.

3.2 Grammar Development and Corpus Annotation

We have been developing grammars for recognizing the above mentioned structural categories 1-3. Our algorithms perform an analysis above the lexical level: they identify structural units of self-narratives on the basis of a combination of lexical, morphological and syntactic features found in the text.
The formulation of such algorithmic descriptions includes two subtasks. The first step is to identify possible linguistic realizations of narrative structural properties and enumerate linguistic markers. The second step is to build a linguistic model of the syntactic behavior of these potential markers.
Accordingly, we started by fishing for potential markers by means of a statistical analysis on the vocabulary extracted from the lemmatized, POS-tagged interview corpus. Words (lemmata, without restrictions on POS category) which occur significantly more frequently in a specific narrative context were added to a list of keywords (potential markers). This yields a list of ~50-60 keywords for each narrative category. Due to the limited size of the corpus, a significant number of units on the statistically extracted keyword list were considered as noise and were deleted. However, many of the most important lexicalized concepts and structural markers were present on the lists: e.g. the keyword list of embedded evaluation contains words related to assessment (*szomorú* 'sad', *rossz* 'bad', *érzés* 'feeling'), words of quotation/narration (*elmond* 'tell', *beszél* 'talk', *hogy* 'that') and adverbs expressing intensity and subjectivity (*kifejezetten,tényleg* 'really'). These lists were then manually enhanced with a set of synonymous expressions, providing a first list of candidate terms, to be added later to a specific semantic dictionary. It is important to note that several elements on the list are ambiguous in that they are relevant to more than one of the above listed narrative structural categories. Besides,

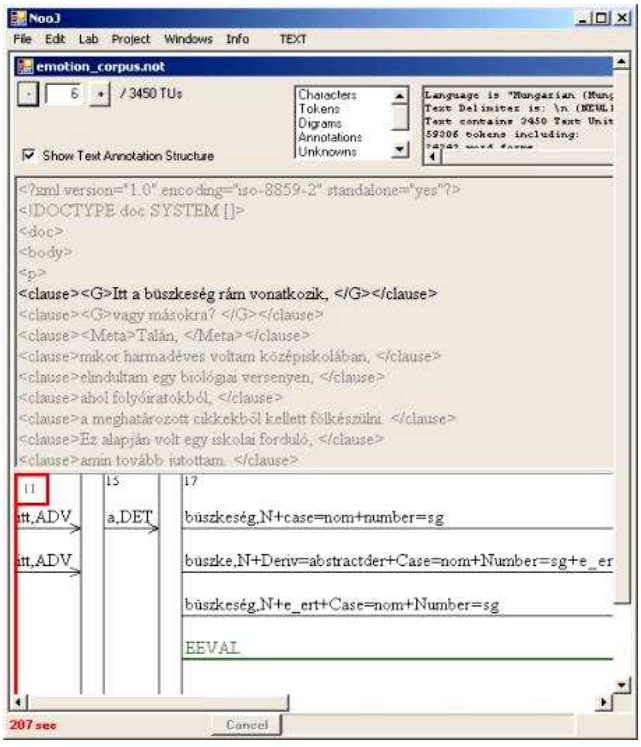

Fig.2.: Annotation of the corpus in NooJ

their unique presence is not sufficient in itself for the narrative clause to be categorized. They only constitute a clue for the elaboration of more detailed linguistic descriptions.
Subsequently, a deep linguistic study of the annotated corpus had to be carried out at the morphological/syntactic level. As a result of this linguistic study, we built up a list of structural markers for each narrative category.

Structural markers correspond to syntactic structures which show a strong correlation with the occurrence of narrative structural categories: according to our hypothesis, they anchor narrative content to linguistic structure. One of the important characteristics of markers is that they do not systematically overlap with constituents in the phrase structure or, in our case, with annotated chunks. This is why it is important that our cascaded syntactic grammars do not replace the text: each level of representation remains accessible at each step of the annotation. We then transform structural markers into grammars which can be applied to the text and yield an annotation which covers the entire scope of the given category, in accordance with the manual annotation at the level of narrative clauses.

Thus, we produced a formal description of the linguistic contexts of lexical/structural markers, realized as a set of NooJ grammars, which can be integrated in our language processing tool chain. The linguistic complexity of our grammars differs considerably among narrative structural categories. Determining the scope of such structural categories requires a linguistic processing which goes beyond lookup-based analysis: unlike most concurrent methods, ours makes extensive use of syntactic analysis. For instance, in the case of evaluation the main difficulty is to identify embedded clauses which express the narrator's view on the events in question. Trigger words can be found in the matrix clause, but the scope of the evaluation extends to every embedded clause which is a syntactic dependent of the matrix clause. For these considerations, structural patterns related to psychological content have to be processed after deep syntactic analysis is completed.

The corpus was annotated on the linguistic level using NooJ's interactive annotation interface. This allows the user to manually select correct hits in a concordance list before annotating the text. Since some of the linguistic anchors of narrative categories cannot be recognized without using pragmatic and other extra-linguistic information, a perfect and totally automated analyzer cannot be conceived at the moment. However, semi-automatic annotation with manual correction allowed us to construct a corpus which can serve as a standard for the elaboration of our automatic rule system.
The grammars developed during the project can also be applied in a non-interactive mode, which makes it possible to annotate new texts fully automatically.

4. Conclusion & Future Work

We have described the structure of a Hungarian corpus for emotion studies. This corpus serves as a resource to test our hypothesis that emotional intensity and narrative structure are related to each other. Therefore, relevant characteristics of narrative structure have been manually annotated. The study of emotional intensity through narrative structure may open a new way for the empirical research of emotional experience.

The second goal of the project is to study linguistic realizations of narrative structural categories. This leads off to developing language processing methods to automatically recognize narrative structural categories and to annotate them directly on the textual level. This annotation was carried out semi-automatically on the corpus, but can be done fully automatically as a part of the Hungarian NLP tool chain in NooJ.

Currently, as a second phase of the project, an English language corpus of the same size is being registered. This corpus will be annotated according to the same guidelines.

5. Acknowledgements

The project was supported by the Hungarian Scientific Research Fund (OTKA /67914).

6. References

Brewer, W.F. & Lichteinstein, E. H. (1982). Stories are to Entertain: A Structural-Affect Theory of Stories. *Journal of Pragmatics*, 6, pp. 473--486.

Bruner, J.S. (1990). *Acts of meaning*. Cambridge: Harvard University Press.

Bucheim, A. & Mergenthaler, E. (2000). The relationship among attachment representation, emotion-abstraction patterns, and narrative style: A computer-based text analysis of the adult attachment interview. *Psychotherapy Research*, 10(4), pp. 390--407.

Gábor, K. (2007). Syntactic Parsing and Named Entity Recognition for Hungarian with Intex. In Koeva, S., Maurel, D. & Silberztein, M (Eds.): *Formaliser les langues avec l'ordinateur: De Intex à Nooj*. Besançon: Presses Universitaires de Franche-Comté, pp. 353--366.

Lambie, J.A. & Marcel, A. J. (2002). Consciousness and the Varieties of Emotion Experience: A Theoretical Framework. *Psychological Review*. 109(2), pp. 219--259.

László, J. (2008). *The Science of Stories*. London, Routledge.

Oatley, K. (1999). Why Fiction May Be Twice as True as Fact: Fiction as Cognitive and Emotional Simulation. *Review of General Psychology*, 3(2), 101--117.

Pólya, T., Kovács, I., Gábor K., & Kabai, P. (2009). Intensity of emotional experience during narration and narrative structure. *Poster presented at the conference of the International Society for Research on Emotion.*

Leuven, p. 131.
Prószéky, G. (1995). Humor: a Morphological System for Corpus Analysis. In Rettig I., Kiefer F., Teubert W. (Eds), *Language Resources for Language Technology*. TELRI, Tihany, Hungary. pp. 149--158.
Sarbin, T. R. (1986). The narrative as a root metaphor for psychology. In T.R. Sarbin (Ed), *Narrative Psychology. The storied nature of human conduct*. New York : Praeger, pp. 3--21.
Scherer, K. R. (2001): Appraisal considered as a process of multilevel sequential checking. In: K. R. Scherer, A. Schorr, T. Johnstone (Eds.): *Appraisal processes in emotion. Theory, methods, research*. Oxford: Oxford University Press. pp. 92–120.
Silberztein, M. (2004). NooJ: A Cooperative, Object-Oriented Architecture for NLP. In Muller C., Royauté J. & Silberztein M. (Eds), *INTEX pour la Linguistique et Traitement Automatique des Langues*. Besançon: Presses Universitaires de Franche-Comté, pp. 359--370.
Silberztein, M. (2008). Complex Annotations with NooJ. In *Proceedings of the 2007 International NooJ Conference*. Newcastle:Cambridge Scholars Publishing, pp. 214--227.
Stein, N. L., & Hernandez, M. V. (2007). Assessing Understanding and Appraisals During Emotional Experience. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment*. Oxford: Oxford University Press, pp. 298--317.
Stern, D.N. (1995). *The Motherhood Constellation*. Basic Books.
Tan, E.S. (1996). *Emotion and the Structure of Narrative Film. Film as an Emotion Machine*. Mahwah: Lawrence Erlbaum Associates.
Váradi, T. 2003. Shallow Parsing of Hungarian Business News. In: *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster: UCREL, Lancaster University. pp. 845--851.
Váradi, T. and Gábor, K. (2004). A magyar Intex fejlesztéséről (On developing the Hungarian Intex module). In *Proceedings of the Second Conference on Hungarian Computational Linguistics.* Szeged University Press, Szeged, pp. 3--10.

Online textual communications annotated with grades of emotion strength

Georgios Paltoglou, Mike Thelwall, Kevan Buckley

School of Computing and IT, University of Wolverhampton, UK
Wulfruna Street, Wolverhampton WV1 1LY, UK
g.paltoglou@wlv.ac.uk, m.thelwall@wlv.ac.uk, k.a.buckley@wlv.ac.uk

Abstract

In this paper, we present two new data sets for textual sentiment analysis. In comparison to other publicly available data sets comprised of product reviews, the new data sets are extracted from social exchanges and debates between people on the web. The first one is a complete crawl of a subsection of the BBC Message Boards spanning four years, where users discuss ethical, religious and news-related issues and the second is a three month complete crawl of the Digg website, one of the most popular social news websites. Human annotators were given the task of judging the level of positive and negative emotion contained on a subset of the extracted data. The aim of the data sets is to provide new standard testbeds for sentiment analysis of textual exchanges between users of online social communities.

1. Introduction

Research in sentiment analysis has mainly focused on data sets extracted from review sites. Prominent examples include the Movie Review data set by Pang et. al (Pang et al., 2002) and general product reviews (Blitzer et al., 2007; Hu and Liu, 2004).

These data sets offer the significant advantage of providing an easily extractable "golden standard" because typically each review is accompanied by a score (i.e. number of stars, thumbs up or thumbs down, etc), which can be easily mapped to a binary (i.e. positive/negative) or multi-point (i.e. one to five stars) score (Pang and Lee, 2005).

Although these data sets have significantly aided the field of textual sentiment analysis by providing standard testbeds for researchers to review and compare their approaches, we believe that they have also limited the field to review-related content.

In this paper, we introduce two new data sets that were created for textual sentiment analysis but are taken from online social communications. The first data set contains discussions extracted from fora and the second contains stories and comments extracted from the Digg[1] website. A subset of the data sets was sampled and given to human annotators to judge the strength of positive and negative emotion contained in them. Both data sets are available from the official CyberEmotions project web site[2].

The following two sections describe the respective data sets and section 4 describes the annotation process. We conclude the paper with some additional comments and conclusions.

2. BBC data set

The British Broadcasting Corporation (BBC) is the world's largest broadcaster, being established under a Royal Charter[3], which has established a framework for scrutinizing the quality, accuracy and impartiality of its reporting (Jowell, 2006). The BBC's web site has a number of publicly-open discussion fora, known as *Message Boards* covering a wide variety of topics[4] that allow registered users to start discussions and post comments on existing discussions. Comments are post-moderated and anything that breaks the "House Rules" is subject to deletion.

Any of the message boards could be have been used for our purposes, but a relatively small number, which have interesting emotional content have been focussed on. Thus, all messages posted on the "Religion and Ethics"[5] and "World / UK News"[6] message boards starting from the launch of the fora (July 2005 and June 2005 respectively) until the beginning of the crawl (June 2009) have been crawled and are made available in the form of a MySql database.

The choice of storing the data in a relational database, in comparison to simpler text files, was made in order to provide researchers with an optimal way of accessing, processing and mining the extracted information. Additionally, information related to the structure of the discussions and the interactions between users would be very difficult to be retained in text files. As a result, all the information that is available on the site, such as discussion threads, comments, quotes, who-replies-to-whom etc. has been successfully captured and is stored in the database, effectively giving researchers the opportunity to completely replicate the content and history of the message boards and allowing them to better study the social interactions between its users. Let it be noted that private information which is not publicly available, such as user e-mails, is also unavailable at the dataset, thus no privacy concerns were raised. The MySQL database schema that was used to store the data is provided in Figure 1.

Some statistics concerning the gathered data are provided below:

- 97,946 separate discussion threads.
- 18,249 users that have posted a comment at least once.
- 2,592,745 distinct comments.

[1] http://www.digg.com

[2] http://www.cyberemotions.eu

[3] http://www.bbc.co.uk/bbctrust/about/how_we_govern/charter_and_agreement/index.shtml

[4] http://www.bbc.co.uk/messageboards/newguide/messageboards_a-z.shtml

[5] http://www.bbc.co.uk/dna/mbreligion/

[6] http://www.bbc.co.uk/dna/mbfivelive/

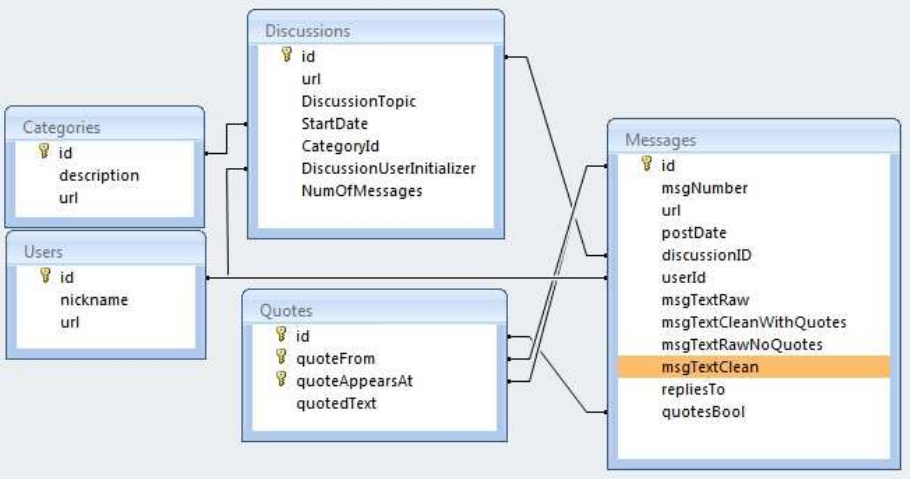

Figure 1: Schema of the database used to store the BBC Message Boards. The content of the posts, stored in table "Messages" is provided in several forms, with and without quotes and HTML markup.

The BBC fora data set was collected using a crawler which for each forum initially downloaded all the forum's header pages containing the title and a variety of summary information about its discussions. The unique discussion ids (later used as a key field in the database) and the URL of the first page of each discussion were then extracted from the header pages. Each discussion was subsequently automatically examined in turn and all of the webpages containing messages were downloaded and collated into a file, resulting in a single file for each discussion. Throughout the process, logs were kept of every transaction, which allowed failed downloads (caused mainly by connection timeouts) to be identified and repeated.

The produced files were then processed in order to remove the HTML annotation, separate the individual posts stored in the file and extract all available information by applying a number of "regular expressions"; for example, the pattern: "`This is a reply to <a.+?\>`" was used in order to extract the post to which a user replies to. This process was chosen because the BBC message boards website is build using a single formatting pattern (i.e. template), which differs only in minor details amongst different subfora. The regular expressions used were general enough to work effectively with no adjustments on the different BBC subfora, but detailed enough in order to extract every available piece of information on the site.

One feature of the dataset that makes it particularly interesting is the structure of discussions, which ensures that when a user submits a comment to a discussion they must specify which existing comment they are replying/responding to. This enables the tracking of conversations/ interactions over time and additionally makes the dataset also useful for research related to network analysis, social interactions etc.

3. Digg data set

Digg[7] is a social news website aiming to help people discover and share content from anywhere on the Internet by submitting stories and subsequently voting and commenting on them. According to Alexa[8] the site ranks as the 96^{th} most popular site on the Internet, higher than other similar sites, such as Slashdot ($1,035^{th}$) or Reddit (310^{th})[9]. One of the site's cornerstone functions is that users are allowed to publicly approve or disapprove submitted stories and comments, processes respectively called "digging" and "burying". The site has stories from a significant variety of subjects, such as politics, entertainment, technology etc. Lastly, in comparison to other similar sites, and to BBC Message Boards, Digg is much more loosely administered making an ideal candidate for collecting un-filtered reactions and emotions as they are expressed online.

The Digg data set was collected making use of the publicly available API of the website[10]. The API allows programmers to access the data that is stored at the website's servers, such as stories, comments, user profiles etc. The quality of the data set is optimal; e.g., the time that a comment is submitted is accurate up the second, in comparison to BBC fora where the time stamp of each post is accurate only to the day.

The data that was gathered from the site is a complete crawl spanning the months February, March and April 2009. The crawl consists of all the stories that were submitted during this period, the comments that they attracted, the diggs (approvals) and buries (disapprovals) that they attained, the users that submitted stories and commented on them and the users that dug stories and comments. The MySQL database schema that was used to store the data is provided in Figure 2. We have tried to keep the BBC and Digg schemas as similar as possible in order to aid research on both data sets, but some changes were necessary in order to store the individual characteristics of both sites (e.g. the "quote" feature of the former and the "digg" feature of the latter).

Some statistics concerning the data that was gathered include:

[7] http://www.digg.com

[8] http://www.alexa.com

[9] Information valid on 24^{th} February 2010.

[10] http://apidoc.digg.com

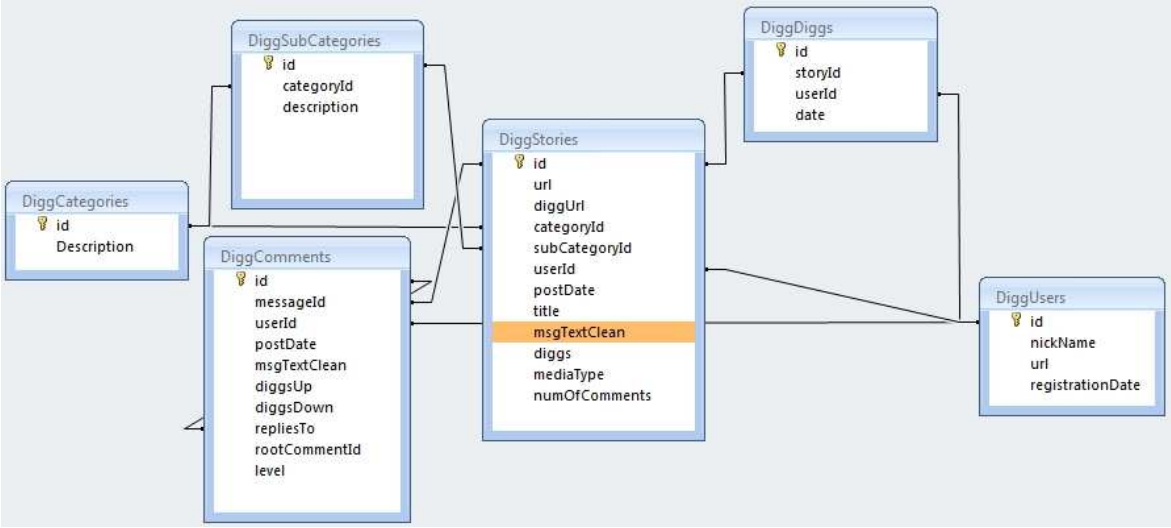

Figure 2: Schema of the database used to store the Digg data.

- 1,195,808 submitted stories.
- 1,646,153 individual comments.
- 877,841 users that submitted a story or comment.
- 14,376,742 approvals (i.e. diggs).

4. Manual Annotation

To enable the manual classification process, a web application was developed that allows participants to log on from anywhere and judge the positive and negative content of the text they are presented with. Figure 3 provides a screenshot of the web application.

In order to obtain reliable human classifications of BBC and Digg comments, two pilot exercises were undertaken with separate samples of data. These were used to identify key classification issues and an appropriate scale. Although there are many ways to measure emotion (Mauss and Robinson, 2009; Wiebe et al., 2005), human classifier subjective judgements were used as an appropriate way to gather sufficient results.

A set of classifier instructions was drafted and refined and the online system constructed to randomly select comments and present them to the classifiers. An extract of the instructions is provided in Appendix A. One of the key outcomes from the pilot exercise was that expressions of energy were difficult to separate from expressions of positive sentiment and so the two were combined. In emotion psychology terminology, this is equating arousal with positive sentiment when it is not associated with negative sentiment (Cornelius, 1996).

Emotions are perceived differently by individuals, partly because of their life experiences and partly because of personality issues (Barrett, 2006) and gender (Stoppard, 1993). In previous work (Pang and Lee, 2005), the ratings from different movie critics were also considered to be incompatible. For the system development, the classifications needed to give a consistent perspective on sentiment in the data, rather than an estimate of the population average perception. As a result, a set of same gender (female) annotators was used and initial testing conducted to identify a homogeneous subset. Five classifiers were initially selected but two were subsequently rejected for giving anomalous results: one gave much higher positive scores than the others and another gave generally inconsistent results.

Table 1: Examples of posts from the BBC forums data set. We also report the positive and negative scores given to the post by the three annotators.

Post Content	Positive Score	Negative Score
Just seen on tv the results of the latest drone missile strike on “tribesmen”.	2	3
It included some stretcher cases being taken into a hospital. Its very	1	4
interesting to look at the footwear of those on the stretchers. Combat boots?	2	4
It’s nice to finally ’meet’ someone with the same idea/thought!!	4 3 4	1 1 1
...what a surprise ! The Sri Lankans say that after the Tsunami, a disaster far far worse than the force 3 hurricane which blew away parts of the American South, they had set up feeding centres within 2 days, there was no looting, no raping, no guns on the streets, no people performing in front of any TV camera they could find....in other words they behaved in a civilised manner. They contrast that with the way the Americans have behaved and ask how they can profess to be a civilised society. A question I have asked many times, but never had a satisfactory answer.	3 1 3	3 4 3

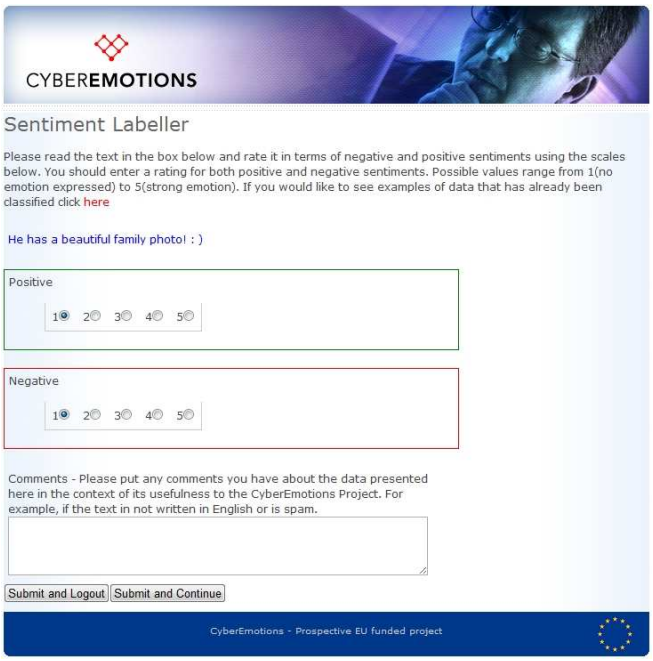

Figure 3: Screenshot of the web application developed for the human annotation process.

For the final classifications, over three thousand comments from each data set were sampled and chosen to be classified by the selected annotators on two 5-point scales for positive and negative sentiment, as follows: `[no positive emotion or energy] 1- 2 - 3 - 4 - 5 [very strong positive emotion]` and `[no negative emotion] 1- 2 - 3 - 4 - 5 [very strong negative emotion]`. The classifiers were given verbal instructions as well as a booklet, motivated by (Wiebe et al., 2005), explaining the task. The booklet also contained a list of emoticons and acronyms with explanations and background context of the task for motivation purposes.

The scales that were used are more fine-grained than most previous studies, which focused on binary or a single 3-star scale. The advantages of the chosen approach is that it allows extrapolation to either binary classification (e.g. a comment that receives a higher positive than negative score by the majority of annotators can be considered as "positive") or to more coarse-grained classification (e.g. scores $2, 3$ can be aggregated into a "somewhat positive/negative emotion" and $4, 5$ can be aggregated into a "very positive/negative emotion").

Finally, it was decided that the annotation process would end as soon as the three annotators had classified at least 1,000 comments in total (i.e. 1,000 posts classified by all three annotators). Therefore, at the end of the annotation phase there were 1,012 BBC posts and 1,084 Digg comments annotated by at least three different people. After a quality check during which we removed 22 problematic BBC posts, which were empty or contained only a url, we finally kept 1,000 BBC posts.

Tables 1 and 2 present some examples of posts that were used in the annotation process, along with the scores that were given to them by the annotators. The first post in both tables has generally been classified as negative by all annotators and the second as positive, showing the generally some agreement is expected between coders. In contrast, the third post has received both high positive and negative scores, indicating that the task is very difficult in some cases and a clear binary classification cannot be always extracted.

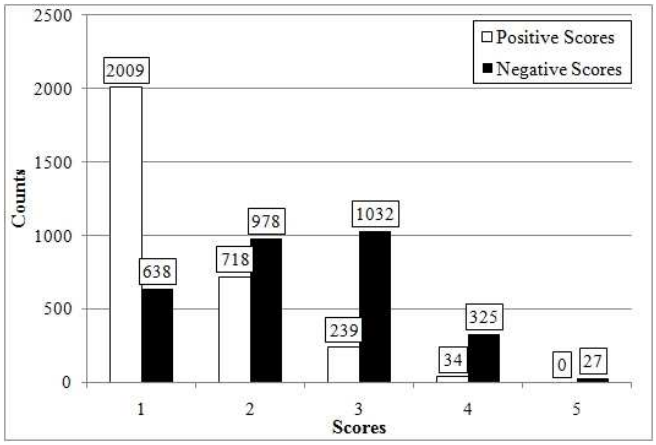

Figure 4: Distribution of Positive and Negative scores given to BBC forum posts by the annotators.

Figures 4 and 5 show the distributions of Positive and Negative scores given to BBC and Digg posts respectively by the human annotators collectively, while figures 6 and 7 show the distributions of scores of each annotator on the BBC data set (respectively, figures 8 and 9 for the Digg data set).

Table 3 reports the inter-agreement between annotators using the Krippendorff alpha coefficient (Krippendorff, 2004) and tables 4 and 5 report the level of agreement between coders individually for the BBC and Digg data sets, respectively. There are no benchmark values for reasonable levels of agreement for such a task but the results are positive enough to indicate that there is a significant degree of agree-

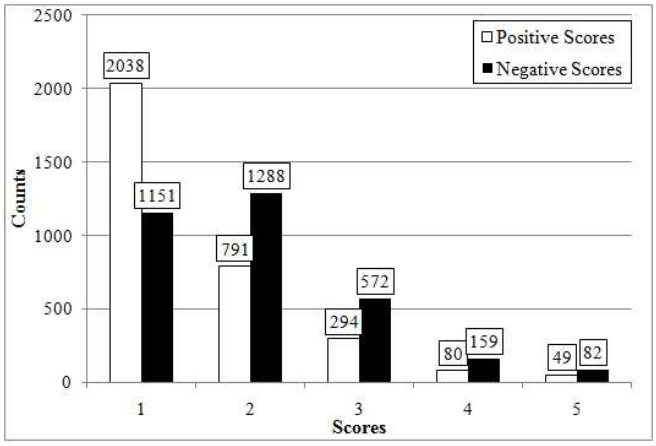

Figure 5: Distribution of Positive and Negative scores given to Digg posts by the annotators.

ment. Additionally, the scores can be compared between each other, despite the fact that the BBC positive set has the lowest overall agreement - probably due to one coder being a little less consistent than the other two.

5. Conclusions

In this paper, we presented two new data sets extracted from BBC Messages Boards and the Digg website. The data sets focus on online communication and debates between users, discussing a diverse set of topics.

A subset of both data sets was sampled and human annotators were used to judge the level of positive and negative emotion contained in the sampled comments.

An interesting quality of the produced data sets is their network structure that has been preserved in their original form.

We believe that the above data sets will prove beneficial in future sentiment analysis research.

Table 2: Example of posts from the Digg data set. We also report the positive and negative scores given to the post by the three annotators.

Post Content	Positive Score	Negative Score
Democracy is dead. Our government killed it and we the people did not revolt and overthrow these tyrants. There is no hope for humanity.	1 1 1	4 3 5
A really nice list of resources!!!	3 5 4	1 1 1
And I think you are a slanderous and ignorant fool, mostly because Rush doesn't use drugs any longer. Great how free speech works, isn't it? I mean, where else can all the people who don't know anything get to talk and then I can show up and show them how ignorant they are? I love it!	3 2 4	3 3 4

Table 3: Table reporting the inter-coder agreement between all three annotators with Krippendorff's multi-coder alpha coefficient.

Data set	Class	Agreement	Weighted Alpha
BBC	positive	50.5%	0.4293
	negative	25.9%	0.5068
Digg	positive	46.0%	0.5010
	negative	29.2%	0.4909

Table 4: Table reporting the inter-coder agreement between all three annotators individually for the BBC data set with Krippendorff's multi-coder alpha coefficient.

Comparison	Class	Agreement	Weighted Alpha
Coder 1 vs. Coder 2	positive	61.9%	0.5493
Coder 1 vs. Coder 3		71.9%	0.6573
Coder 2 vs. Coder 3		59.9%	0.5218
Coder 1 vs. Coder 2	negative	40.0%	0.5773
Coder 1 vs. Coder 3		50.7%	0.6387
Coder 2 vs. Coder 3		48.3%	0.6601

6. Acknowledgements

This work was supported by a European Union grant by the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT. It is part of the CyberEmotions project (contract 231323).

7. References

L. F. Barrett. 2006. Valence is a basic building block of emotional life. *Journal of Research in Personality*, 40(1):35–55.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*. The Association for Computer Linguistics.

R. R. Cornelius. 1996. *The science of emotion*. Upper Saddle River, NJ: Prentice Hall.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*, pages 755–760. AAAI Press / The MIT Press.

Table 5: Table reporting the inter-coder agreement between all three annotators individually for the Digg data set with Krippendorff's multi-coder alpha coefficient.

Comparison	Class	Agreement	Weighted Alpha
Coder 1 vs. Coder 2	positive	61.2%	0.5839
Coder 1 vs. Coder 3		65.9%	0.6787
Coder 2 vs. Coder 3		57.0%	0.6121
Coder 1 vs. Coder 2	negative	44.2%	0.5651
Coder 1 vs. Coder 3		59.7%	0.6756
Coder 2 vs. Coder 3		41.2%	0.6129

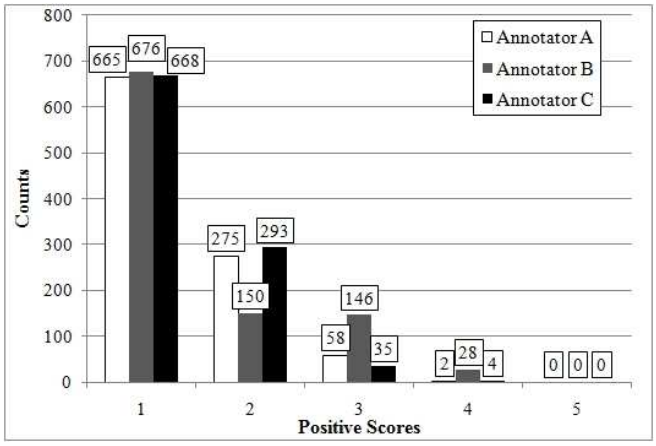

Figure 6: Distribution of Positive scores given to BBC forum posts by each annotator.

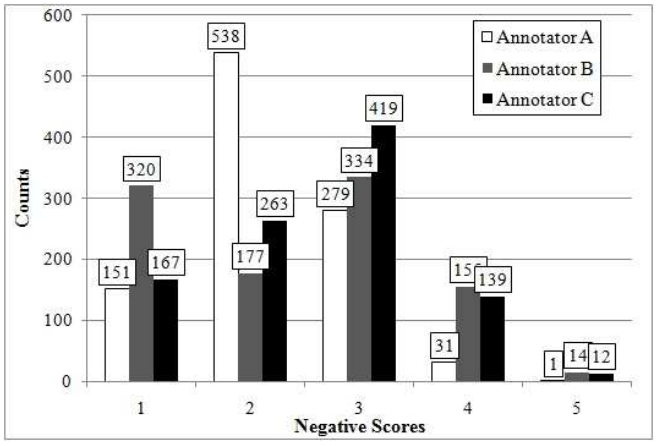

Figure 7: Distribution of Negative scores given to BBC forum posts by each annotator.

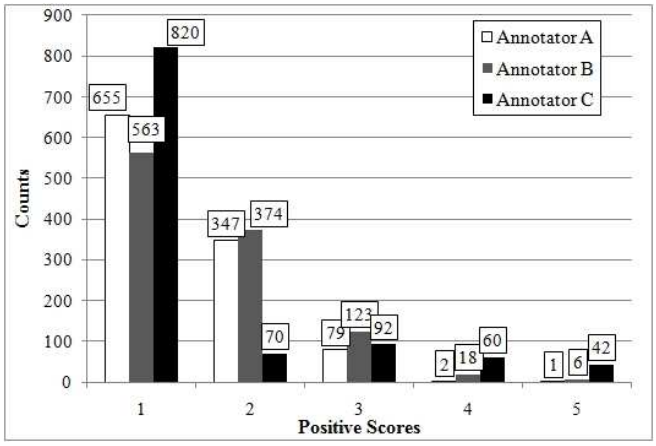

Figure 8: Distribution of Positive scores given to Digg posts by each annotator.

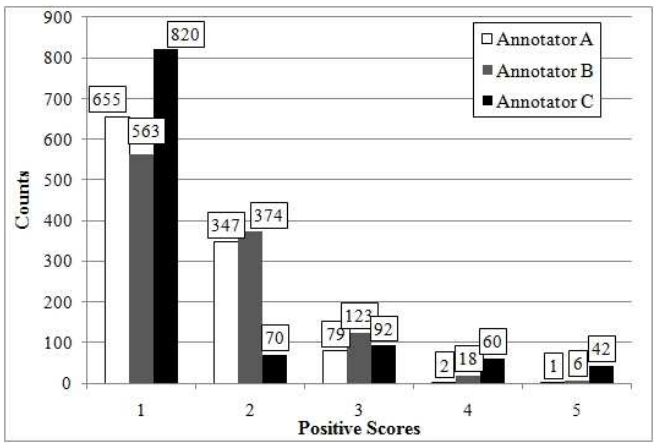

Figure 9: Distribution of Negative scores given to Digg posts by each annotator.

Tessa Jowell. 2006. A public service for all: the bbc in the digital age. White paper, Department for Culture, Media and Sport, March. Available online (76 pages).

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage.

Iris B. Mauss and Michael D. Robinson. 2009. Measures of emotion: A review. *Cognition & Emotion*, 23(2):209–237.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, Morristown, NJ, USA. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Gruchy C. D. G. Stoppard, J. M. 1993. Valence is a basic building block of emotional life. *Personality and Social Psychology Bulletin*, 19(2):143 – 150.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0.

Appendix:

A Classifier Instructions (extract)

Below are descriptions of the judgements for you to make. There are no formal criteria for these judgements. We don't know formal criteria for these judgements! We want you to use your human knowledge and intuition to make your decisions.

You will see a set of comments extracted from the public comment sections of BBC Forums / Digg. Please classify each comment for (a) the strength of positive emotion that it contains and (b) the strength of negative emotion that it contains. Please treat the two emotions as completely separate and don't let one emotion cancel out the other emotion to any extent.

Judge the comments from the perspective of the content of the text, not the author's emotional state or the intended reader's likely emotional state. In other words, the question that you are asking for each comment is: "What emotion is coded inside the text?" Ignore whether the emotion is targeted at someone or not, just identify whether emotion is present in any form.

Code each comment for the degree to which it expresses positive emotion or energy. Excitement, enthusiasm or energy should be counted as positive emotion here. If you think that the punctuation emphasizes the positive emotion or energy in any way then include this in your rating. The scale for positive emotion or energy is:

> [no positive emotion or energy] 1- 2 - 3 - 4 - 5 [very strong positive emotion]

- Allocate 1 if the comment contains no positive emotion or energy.
- Allocate 5 if the comment contains very strong positive emotion.
- Allocate a number between 2 and 4 if the comment contains some positive emotion but not very strong positive emotion. Use your judgement about the exact positive emotion strength.

Code each comment for the degree to which it expresses negative emotion or is negative. If you think that the punctuation emphasizes the negative emotion in any way then include this in your rating. The scale for negative emotion is:

> [no negative emotion] 1- 2 - 3 - 4 - 5 [very strong negative emotion]

- Allocate 1 if the comment contains no negative emotion at all.
- Allocate 5 if the comment contains very strong negative emotion.
- Allocate a number between 2 and 4 if the comment contains some negative emotion but not very strong negative emotion. Use your judgement about the exact negative emotion strength.

When making judgements, please be as consistent with your previous decisions as possible. Also, please interpret emotion within the individual comment that it appears and ignore all other comments.

Multimodal Russian Corpus (MURCO): Studying Emotions

Elena Grishina
Institute of Russian Language RAS
18/2 Volkhonka st., Moscow, Russia
rudi2007@yandex.ru

Abstract

The paper introduces the Multimodal Russian Corpus (MURCO), which has been created in the framework of the Russian National Corpus (RNC). The MURCO provides the users with the great amount of phonetic, orthoepic, intonational information related to Russian. Moreover, the deeply annotated part of the MURCO (DA-MURCO) contains the data concerning Russian gesticulation, speech act system, types of vocal gestures and interjections in Russian, and may be used to study affections and emotions. The access to The Corpus is free. The paper describes the main types of data concerning affects and emotions, which may be obtained from the DA-MURCO in spite the fact that this corpus is not intended for the direct access to the emotion data.

1. What is MURCO?

The Multimodal Russian Corpus (MURCO) is the new project in the framework of the Russian National Corpus (RNC, www.ruscorpora.ru). The pilot version of the MURCO has been open for general access since April, 2010.

Since the project was described in my papers (Grishina, 2009a; Grishina, 2009b; Grishina, 2010a), I don't intend to describe the MURCO at great length. I am planning to outline the Corpus and to present its perspectives in relation to the study of emotions.

The MURCO is the result of the further development of the Spoken Sobcorpus of the RNC. The Spoken Subcorpus includes circa 8.5 million tokens and contains the texts of 3 types: public spoken Russian, private spoken Russian, and movie speech (the volume of the last is circa 4 million tokens).

The Spoken Subcorpus does not include the oral speech proper; it includes only transcripts of the spoken texts (Grishina 2006). To improve it and to supplement its searching capacity, we have decided to supply the transcripts with the sound and video tracks. To avoid the problem related to the copyright offence and the privacy invasion we have used the cinematographic material in the MURCO.

Naturally, in the future we are also going to include in the MURCO the patterns of the public and private spoken Russian, but the cinematographic Russian is the most appropriate material to begin the project with. It should be mentioned inter alia that the usage of the cinematographic material to elaborate and test the annotation system of the pioneering corpus is far more promising than the usage of the "natural" (public or private) spoken Russian. The main reason for it is the fact that movies include exceptionally manifold set of situations, and this situational variety results in the linguistic variety. Therefore, to annotate the movie Russian we need greater number of definitions and more elaborated system of concepts than to annotate the "real-life" Russian. In other words, the annotation of the cinematographic Russian will be useful for the mark-up of the "natural" Russian, but the opposite is not correct.

The MURCO is the collection of the clixts. A ***clixt*** is the pair of a *clip* and the corresponding *text* (i.e. the corresponding part of a movie transcript). It is supposed that a user will have the opportunity to download not only the text component of a clixt (=marked up transcript), but also its sound/video component, so after downloading a user may employ any program to analyze it. The duration of a clip is within the interval of 5-20 sec.

As we have mentioned above, the total volume of the movie subcorpus is about 4 million tokens. This token volume corresponds to circa 300 hours of sound- and video track. Therefore, being fulfilled the MURCO presents one of the largest open multimodal sources.

2. Data Validity

The first question which occurs when we discuss the usage of the cinematographic data in linguistic researches is the question of their validity. The standard attitude to the movie speech considers it to be "written to be spoken" and "artificial" (Sinclair, 2004). The closer look at the problem, however, shows that the discrepancies between the spontaneous "real-life" spoken language and the movie speech are not crucial and deal mainly with the higher/lower degree of text coherence. The other spoken features, first of all the set and the structure of discourse markers (lexical, morphological, and syntactical), the intonation patterns and the accompanying gesticulation are similar in the cinematographic and real face-to-face communications (Grishina, 2007a; Grishina, 2007b; Forchini, 2009).

As for the emotion studies, it is well known that one of the generally accepted ways of data acquisition here is to get the actors who are supposed to perform the desired mise-en-scènes involved in the investigation process. In a very instructive paper (Busso & Narayanan, 2008) the authors have shown distinctly that the reason of the main drawbacks concerning this way of the data accessing roots is the wrong methodology. As for the authors' opinion, two main shortcomings are as follows: the lack of acting technique and the lack of contextualization of the uttered phases. It is absolutely obvious that the cinematographic data are free from these drawbacks.

3. Types of Annotation in MURCO

We should emphasize the fact that the MURCO is not

intended for studying emotions and affections: it has been designed as the specialized resource oriented basically at standard linguistic researches. However, this resource may be used efficiently when we study emotions and affections in the light of their linguistic manifestation. In my paper I will try to number the ways to obtain emotional information from the MURCO.

3.1 Via Word to Emotion

The MURCO is marked up from different points of view (Grishina, 2010a). Some types of annotation are standard RNC types of annotation (metatextual, morphological, semantic annotation), some types are special for the spoken component of the RNC and, naturally, are preserved in the MURCO (sociological and accentological annotation), and, finally, some of the mark-up dimensions are specific only for the MURCO (orthoepic, speech act and gesture annotation). We do not plan to describe all these types of annotation in detail (it has been done in our early papers); we only want to illustrate the annotation zones, which can be useful when we study emotions and affections.

It is obvious that the simplest way to obtain the emotion data from the MURCO is the word entry to the Corpus. We can find the data using the special morphological form of a word. For example, we may form the following query: "verb *nenavidet'* (to hate), 1st person, Present, Singular", i.e. 'I hate smb or smth'. This query gives us the possibility to obtain clixts, in which different characters express their hatred of smd/smth in actual mode.

To widen the scope of the data we can also use the set of synonyms, for example, "*otvratitel'nyj* 'disgusting', *gadkij* 'nasty', *gnusnyj* 'abominable', *merzkij* 'loathsome', *omerzitel'nyj* 'sickening', *pakostnyj* 'mean, foul'", and some others. This synonymic row lets us receive all possible types of pronouncing of these adjectives, so we can analyze the phonetic, intonation and gesture characteristics of the adjectives of disgust in Russian. Moreover, the morphological annotation in the MURCO makes it possible to distinguish the usage of these adjectives as attributes and as nominal predicates; this distinction is very important in Russian. It seems that the adjectives-attributes are used to describe the speaker's emotions, while the adjectives-predicates express them directly.

Finally, to collect clixts with the preselected emotions we may form the request which consists of emotionally tinged syntactic constructions. The example of such constructions in Russian is the word-combination *chto za* X_{nom}*!* 'What X!' This construction may express both high appraisal and condemnation (so, *Chto za pogoda!* 'What weather!' may mean both 'good weather' and 'bad one'). Obviously, it may be very useful for any investigator of emotions to study the discrepancies between the opposite emotional colors of the same syntactic construction.

It should be mentioned in conclusion that the MURCO annotation makes it possible to form the subcorpora of 1) the masculine and feminine cues; 2) the cues which belong to the actors of certain age; 3) the cues which were pronounced by certain actor. Therefore, we may try to arrive to the conclusions concerning sociology of emotions.

3.2 Via Speech Act to Emotion

The clixts in the MURCO receive the annotation of 2 types: 1) obligatory annotation (metatextual, morphological, semantic, sociological, accentological, orthoepic), and 2) optional annotation (speech act and gesture). The last one is optional by reason of its labour-intensiveness and impossibility to automate the mark-up process. Therefore, the clixts, which are annotated from the both points of view, form the subset of the MURCO. This subcollection is named 'the deeply annotated subcorpus' (DA-MURCO). The volume of the DA-MURCO is supposed to be circa 0.5 million tokens, or 40 hours of phonation.

The speech act annotation in the DA-MURCO covers 5 thematic areas: the mark-up of 1) speech acts proper, 2) the types of repetitions, 3) the speech manner, and 4) the types of vocal gestures, non-verbal words and interjections.

3.2.1. Speech Act Annotation

The list of the Russian speech acts includes about 150 items, grouped into 13 types: *Address or call, Agreement, Assertion, Citation, Complimentary, Critical utterance, Etiquette formula, Imperative, Joke, Modal utterance or performative, Negation, Question, Trade utterance.* Most of these types are ambivalent from the point of view of emotional characteristics, that is they include both emotionally colored speech acts (e.g. *to* $monkey_{emotion}$ and *to* $quote_{neutral}$ in the type *Citation*, *to ask smb to do* $smth_{neutral}$, *to* $advise_{neutral}$, *to* $order_{neutral}$ and *to* $demand_{emotion}$ or to $supplicate_{emotion}$ in the type *Imperative*, and so on). Some types, however, are characterized with the inherent emotionality, that is all speech acts, which are affiliated to these types, are emotionally colored (*Critical Utterances, Complimentary, Joke*). Naturally, to study emotions both emotionally colored types of speech acts and ambivalent ones are useful. As for the emotionally colored types, it is quite obvious. In respect to the ambivalent types of speech acts, it is very interesting, for example, to define whether the neutral and the emotional speech acts preserve their neutrality and emotionality in different types of situations or not.

Therefore, the speech act annotation makes it possible to request clixts, which contain this or that type of speech acts, no matter what their lexical, morphological and syntactic structures are.

3.2.2. Repetitions

It is well known that the repetitions are of great importance in the spoken speech and are directly connected with the expression of emotions. Therefore, various types of repetitions, which are marked up in the DA-MURCO, may become very useful for the researchers of emotions.

Three types of repetitions are marked up in the DA-MURCO. Firstly, we may obtain the clixts which contain the repetitions of different **structure** and try to analyze their possible connection with the emotionality.

The structural types of repetitions are as follows:
a) one-word vs. many-word repetitions
b) envelope repetitions:

Beregi	*ruku*	*Sen'a*	***beregi***
Be careful	*with your hand*	*Sen'a*	***be careful***

c) baton repetitions

Ja ne trus	*no ja* ***bojus'.***	***Bojus'***	*smogu li ja*
I am not coward	*but* ***I'm afraid.***	***I'm afraid***	*I can't do it*

d) permutation repetitions

Tak	*chto*	***delat'***	***budem?***	*Chto*	***budem***	***delat'?***
Well	*what*	***to do***	***we ought?***	*What*	***we ought***	***to do?***

Well, what we ought to do?

Secondly, we may request the clixts which include the repetitions of different **communicative** structure, i.e. the repetitions which are distributed between the participants of a communicative act:
a) echo repetitions

1st: *On hochet*	*chtoby ty sygral*	***Lenina***	*v jego spektakle.*
1st: *He wants*	*you to play*	***Lenin***	*in his performance.*
2nd:	*Gad!*	***Lenina!***	
2nd:	*Son of a bitch!*	***Lenin!***	

b) change of addressee

(to 1st) *Vot usy vam*	***vylityj***	*Volod'ka!*	(to 2nd) ***Vylityj!***
(to 1st) *If you have the moustache*	*you'll be* ***the spitting image*** *of*	*Volod'ka!*	(to 2nd) ***The spitting image!***

c) forwarding repetition

1st (to audience) *Pobeditel'nicej stala…*	2nd (to 1st) ***Minutochku.***	1st (to audience) ***Minutochku!***
1st (to audience) *The winner is…*	2nd (to 1st) ***Just a minute.***	1st (to audience) ***Just a minute!***

d) overinterrogations

1st: *Eto kajuta*	***shestnadcataja****?*	2nd: ***Shestnadcataja.***	
1st: *Is it the cabin*	*number* ***sixteen****?*	2nd: ***Sixteen****.*	
1st: *Gde*	***etot*** *bol'noj?*	2nd: ***Etot****?*	*Tam.*
1st: *Where is*	***this*** *patient?*	2nd: ***This*** *one?*	*There.*

And, finally, the repetitions may be of different **intensity** and **emotional color**.
a) single vs. multiple repetitions
b) repetitions with intensifiers:

Blagodarim vas	*za vashu* ***interesnuju***	*o c h e n '* ***interesnuju***	*lekciju*
Thank you	*for the* ***interesting***	*v e r y* ***interesting***	*lecture*

c) monkey repetitions:

Nu vot	*voshla*	*kak rodnaja.*	*a ty:*	***"ne vojd'ot, ne vojd'ot!"***
Here we are	*it's gone in*	*OK.*	*You was wrong saying:*	***"it will hardly go in! it will hardly go in!"***

Apparently, the consistent description and manifold set of the repetitions appear to be very useful not only while studying the specificity of the spoken Russian, but also in emotion studies.

3.2.3. Speech Manner

The speech manner annotation in the DA-MURCO describes the speech acts from the point of view of 1) speech tempo (quick speech, chanting/scanning, declamation, speech with extra pauses), 2) speech volume (whisper, loud shout, muffled shout), 3) emotional color (crying, laughing). Obviously, the last two positions are promising from the point of view of emotion studies.

3.2.4. Non-Verbal Words and Interjections

It is quite evident that non-verbal words and interjections are the most striking and direct illustrations of the emotional state of a speaker. Therefore, it is quite possible to obtain the information concerning emotions from the MURCO just making the query which contains this or that interjection or non-verbal word. The problem, however, lies in the fact that non-verbal words and interjections are polysemic, so the direct query *Ah* or *Oh* will give us the great number of varied contexts and we will be forced to do a big chunk of work to distinguish 1) emotionally colored and neutral interjections and 2) interjections expressing different emotions.
The recent investigation of the Russian vocal gestures *Ah* and *Oh* (Grishina, 2009c; Grishina, 2010b) shows that these units have three types of usage.

1) *Ah/Oh* as **exclamations**. The exclamation *Oh* means 'suppressed pain', 'smth unpleasant', 'exercise stress', 'intensity of feelings'; *Oh* is characterized with descending tone and throaty phonation. The exclamation *Ah* means 'uninhibited pain', 'unexpectedness', 'fright', 'horror'; *Ah* is also characterized with descending tone and throaty phonation, but sometimes the tone of the *Ah* in the sense 'horror' is high and static.
2) *Ah/Oh* as **interjections**. The interjection *Oh* means basically 'surprise' and has also some derived meanings, e.g. 'high appraisal', 'sneer', and others. The interjection *Ah* means basically 'realization' or 'comprehension' and has also some derived meanings, e.g. 'intellectual satisfaction', 'recognition.' Both *Ah* and *Oh* are characterized with the ascending-descending, or undulatory tone.
3) *Ah/Oh* as **particles**. The particle *Oh* is the deictic one and means 'indication', 'object fixing', and some others. The particle *Oh* is characterized with even ascending or even descending tone, and also with the glottal stop at the beginning of its phonation. The particle *Ah* has 2 main meanings: a) *Ah* as a interrogative particle basically means 'question', 'echo-question', 'answer to address' and is characterized with the ascending tone; b) *Ah* as a negative particle means 'disregard' 'annoyance' and is characterized with the descending tone.

So, we can see that the straightforward lexical query of non-verbal words and interjections gives us too heterogeneous data to deal with. Therefore, in cases when we plan to investigate relatively frequent phenomena of the

kind it would be more convenient to use the DA-MURCO, where all non-verbal words and interjections are described from the point of view of their contextual meaning.

3.2 Via Gesture to Emotion

It is the wide-spread opinion that the gesticulation (including the facial expressions) along with the intonation and the phonetic indicators are the main and principal media to convey the emotional information. The MURCO seems to be the resource which is generally accessible and quite considerable in terms of its volume; moreover, the MURCO includes a lot of video tracks. So, the existing gesture annotation in the DA-MURCO (see in detail (Grishina, 2010)) ought to supply a user with the shortest ways to the required emotion information obtained by means of gesticulation. It should be specially mentioned that the gesture annotation makes it possible to investigate the emotions of the silent participants of a communication act, in contrast to the other linguistic units (phonetic, intonation, lexical, grammatical).

Any gesture in the DA-MURCO is supplied with 2 types of characteristics: 1) objective, which describes a gesture from the point of view of active/passive organs, their orientation relative to the speaker's body and their movement directions; 2) subjective. The last ones are the triads of gesture type, gesture meaning, and gesture name. Till the moment we have marked out about 250 gesture meanings, which are grouped into 14 gesture types. To designate these gesture meanings more than 400 gesture names are used. These gesture names are the natural Russian words and word combinations, which describe Russian gesticulation and facial expressions.

The gesture types are as follows: *Adopted, Conventional, Corporate, Critical, Decorative, Deictic, Etiquette gestures, Gestures – speech acts, Gestures of inner state, Iconic, Physiological, Regulating, Rhetorical, Searching gestures.*

Among these 14 types of gestures the gestures of inner state and critical gestures are directly connected with the emotion studies, so we would like to sketch them in.

The group of the **gestures of inner state** includes the following gesture meaning[1]:

Gesture meaning	Gesture name
1) admiration	move back from smth, sidelong glance
2) affectation	stick one's little finger out
3) affection	stroke smb's head; embrace smb
4) alienation	shrug one's shoulders; lift one's hands
5) anticipation	lick one's lips
6) anxiety	one's hand to one's lips
7) apprehension	glance back
8) archness	lift one's eyebrow; sidelong glance
9) boredom	beat a tattoo; lean one's head on one's arm
10) caution	shield oneself; glance back
11) chagrin	shake one's head; lean one's head on one's arm
12) concentration	close one's eyes; lick one's lips
13) confidence	nod; cross one's legs
14) coquetry	crinkle one's nose; blow kisses to smb
15) defiance	brush aside; screw up one's face
16) despair	grasp one's head; strike one's hand on smth
17) disappointment	lift one's eyebrows; press one's lips together
18) disgust	screw up one's face
19) displeasure	turn away; throw smth off
20) distrust	shake one's head; blink
21) embarrassment	close one's eyes; shake one's head; turn away
22) enervation	toss smth up; lean against a wall
23) friendliness	stroke smb's shoulder; embrace smb
24) fright	close one's eyes tight; recoil
25) frolic	throw back one's head; give smb a flick on the nose
26) grief	grasp one's head; bite one's lips
27) guess	cover one's mouth; grasp one's head
28) high appraisal	lift one's eyebrows; shake one's head; throw up one's hands
29) impatience	stamp one's foot
30) indignation	throw up one's hands; open one's eyes wide
31) interest	lift one's eyebrow; bend forward
32) irritability	roll up one's eyes; close one's eyes
33) joy	embrace smb; clap one's hands
34) lust	eye smb from head to foot; close one's eyes
35) meditation	beat a tattoo; bite one's lips
36) nervousness	with one's hands to one's heart; beat a tattoo
37) perplexity	look round the interlocutors; touch smth
38) pride	throw back one's head; stalk along
39) relief	close one's eyes
40) resolution	pull one's hat over one's eyes; roll up one's sleeves
41) sadness	bend one's head on one's breast
42) satisfaction	nod; hang of the head on one side
43) seducing	touch smb; stroke one's throat
44) shame	close one's eyes; close one's face with one's hands
45) shock	look round the interlocutors; open one's eyes wide
46) solidarity	take smb by the hand; wink towards smb
47) sudden recollection	cover one's mouth; knock one's head; dive a jump
48) surprise	clasp one's hands; open one's mouth
49) veneration	kneel
50) waiting	lean against smth; to fold one' s arms
51) willingness	straighten one's tie; rub one's hands

Table 1.

The group of the **critical gestures** includes the following

[1] Every gesture meaning is supplied in the Tables 1-2 with one or two gesture names. Naturally, we have no possibilities to publish the complete list of gestures.

gesture meaning:

Gesture meaning	Gesture name
1) you are a fool!	give the screw-loose sign; shake one's hand
2) criticism	shake one's finger; shake one's head
3) mocking	shake one's head from side to side
4) who cares!	spread out one's arms

Table 2.

We can see that the gesture annotation in the DA-MURCO gives a user the data for the multidirectional emotion researches. Firstly, we can study a certain gesture as a medium of different emotions (e.g. the gesture *to beat a tattoo* may express *boredom, meditation, nervousness, alienation, waiting,* and *the search for necessary word* – one of the searching gestures) and try to define the significant discrepancies between different ways of realization of the same gesture in different emotional situations.

On the other hand, we can analyze the cluster of different gestures which correlate with the same emotion. For example, *an embarrassment* may be expressed, among others, by means of the following gestures: *to hunch, to close one's eyes, to turn away from smb, to look away, to drop one's eyes, to shade one's face with one's hand, to cover one's mouth, to bend one's head.* All these gestures have the same dominant: a confused person tries to attract a little attention, to become unobtrusive, and to achieve this goal he leaves the communicative zone and breaks the visual and speech contact with his interlocutor.

In conclusion, I'd like to mention that we can combine the speech act queries with the gestures ones, and this combination lets us obtain the clixts which may contain not only required gesture and required emotion, but also this or that speech act, interjection, non-verbal word, type of repetitions and speech manner. This potential extends our ability to study emotions to a considerable degree. And if we keep in mind the relative universality of emotions and affects, it will be possible to take advantage of the MURCO as a whole and of the DA-MURCO particularly not only by Russian speakers, but also by the users who do not speak Russian.

4. Acknowledgements

The work of the MURCO group is supported by the program "Genesis and Interaction of Social, Cultural and Language Communities" of the Russian Academy of Sciences. The author's investigation is supported by the 1) RFBR[2] (RFFI) under the grant 08-06-00371a and the grant "Elaboration of Multimodal Russian Corpus (MURCO) within the framework of Russian National Corpus (www.ruscorpora.ru)", and 2) RFH[3] (RGNF) under the grants "Russian Gesticulation according to the Cinematographic Data" and "Deeply Annotated Multimodal Russian Corpus: Elaboration and Creation".

5. References

Busso, C. , Narayanan, S. (2008). Recording Audio-Visual Emotional Databases from Actors : a Closer Look. In *6th International Conference on Language Resources and Evaluation.* Marrakesh: ELRA.

Forchini, P. (2009). Spontaneity reloaded: American face-to-face and movie conversation compared. In *Corpus Linguistics 2009. Abstracts The 5th Corpus Linguistics Conference, 20-23 July 2009, Liverpool, p. 118.*

Grishina, E. (2006). Spoken Russian in the Russian National Corpus (RNC). In *LREC'2006: 5th International Conference on Language Resources and Evaluation. ELRA*, pp. 121-124.

Grishina, E. (2007a). O markerah razgovornoj rechi (predvaritel'noje issledovanije podkorpusa kino v Nacional'nom korpuse russkogo jazyka). In *Kompjuternaja lingvistika i intellektual'nyje tehnologii. Trudy mezhdunarodnoj konferencii "Dialog 2007"*. Moscow, RSGU, pp. 147-156.

Grishina, E. (2007b). Text Navigators in Spoken Russian. In *Proceedings of the workshop "Representation of Semantic Structure of Spoken Speech"* (*CAEPIA'2007, Spain, 2007, 12-16.11.07, Salamanca*). Salamanca, pp. 39-50.

Grishina, E. (2009a). Multimodal Russian Corpus (MURCO): Types of annotation and annotator's workbenches. In *Corpus Linguistics'2009*. Liverpool (forthcoming).

Grishina, E. (2009b). Multimodal Russian Corpus (MURCO): general structure and user interface. In *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference. Smolenice, Slovakia, 25-27 November 2009. Proceedings.* Brno, Tribun, pp. 119-131.

Grishina, E. (2009c). K voprosu o sootnoshenii slova i zhesta: vokal'nyj zhest Oh v ustnoj rechi. In *Komp'juternaja lingvistika i intellektual'nyje tekhnologii (Mezhdunarodnaja konferencija "Dialog 2009", 8(15)). Computational Linguistics and Intellectual Technologies (Annual International Conference "Dialogue 2009", 8(15))*. Moscow: RSGU, pp. 80-90.

Grishina, E. (2010a). Multimodal Russian Corpus (MURCO): First Steps. In *LREC'2010* (forthcoming).

Grishina, E. (2010b). Vokal'nyj zhest Oh v ustnoj rechi. In *Komp'juternaja lingvistika i intellektual'nyje technologii (Mezhdunarodnaja konferencija "Dialog 2010", 9(16)). Computational Linguistics and Intellectual Technologies (Annual International Conference "Dialogue 2010", 9(16))*. Moscow: RSGU, (forthcoming).

Sinclair, J.M. (2004). Corpus Creation. In Corpus Linguistics: Readings in a Widening Discipline. London, Continuum, pp. 78-84.

[2] The Russian Fund of Basic Researches.
[3] The Russian Fund of Humanity.

The Rovereto Emotion and Cooperation Corpus

Federica Cavicchio, Massimo Poesio
Mind and Brain Center/ Corso Bettini 31, 38068 Rovereto (Tn) Italy
Department of Information Engineering and Computer Science / Via Sommarive 14, 38123 Povo (Tn) Italy
federica.cavicchio@unitn.it, massimo.poesio@unitn.it

Abstract

The Rovereto Emotion and Cooperation Corpus (RECC) is a new resource collected to investigate the relationship between emotions and language in an interactive setting. The coding and decoding of emotions in face to face interactions belong to an area of language studies -often called paralinguistics- where still many unsolved questions are present. In this paper we present a new corpus with a focus on emotion and cooperation relationship. Beside linguistics aspects of interaction, we take into account other aspects of communication such as pragmatics, facial expressions and gaze direction. Many corpora have been collected and annotated in the last years to study emotions. Many of these corpora have been collected in a natural setting and sometimes the resulting data are quite wild and difficult to classify and analyse. As a consequence, coding schemes issued to analyse emotions have been shown not to be entirely reliable. Because of those previous results, we collected a corpus in which emotions are pointed out by psychophysiological indexes (such as ElectroCardioGram and Galvanic Skin Conductance) which are highly reliable. RECC corpus is up to now a one of a kind resource allowing the study of linguistics, pragmatics, cognitive and behavioral aspects of cooperation and emotions.

1. Introduction

In the last years many multimodal corpora have been collected. These corpora have been recorded in several languages and have being elicited with different methodologies. Among the goals of these corpora there is shading light on crucial aspects of speech production. Some of the main research questions are how language and gesture correlate with each other (Kipp et al., 2006) and how emotion expression modifies speech (Magno Caldognetto et al., 2004) and gesture (Poggi & Vincze, 2008). On the other hand, multimodal coding schemes are mainly focused on dialogue acts, topic segmentation and the so called "emotional" or "social" area of communication (Carletta, 2007; Pianesi et al., 2006).

The large use of corpora in linguistics and engineering has raised questions on coding scheme reliability. The aim of testing coding scheme reliability is to assess whether a scheme is able to capture observable reality and eventually allows some generalizations. From mid Nineties, the Kappa statistic has begun to be applied to validate coding scheme reliability. Basically, the Kappa statistic is a statistical method to assess agreement among a group of observers. Kappa has been used to validate some multimodal coding schemes too. However, up to now many coding schemes to codify emotions have a very low agreement among annotators (Reidsma & Carletta, 2008; Douglas-Cowie et al., 2005; Pianesi et al., 2006; Allwood et al., 2006, 2007). This could be due to the nature of emotion data. In fact, annotation of mental and emotional states of mind is a very demanding task. The low annotation agreement which affects multimodal corpora validation could also be due to the nature of the Kappa statistics. The assumption underlining the use of Kappa as a reliability measure is that coding scheme categories are mutually exclusive and equally distinct one another. This is clearly difficult to be obtained in multimodal corpora annotation, as communication channels (i.e. voice, face movements, gestures and posture) are deeply interconnected one another (Cavicchio & Poesio, 2009).

The emotion annotation methods followed so far can be gathered in two groups. The first group is represented by Craggs and Woods' work (2004). In Craggs and Woods' opinion, annotators must label the given emotion with a main emotive term (e. g. anger, sadness, joy etc.) correcting the emotional state with a score ranging from 1 (low) to 5 (very high). For the second group, a three steps rank scale based on emotion valence -positive, neutral and negative (Martin et al. 2006; Callejas et al., 2008; Devillers et al., 2005)- is used to annotate a variety of corpora mostly recorded in natural settings. But both these methods had quite poor results in terms of annotators' agreement.

Keeping this in mind, we do not label emotions directly but we indirectly attribute arousal and valence values. According to the appraisal theory of emotion, an emotion affects the autonomic nervous system (psychophysiological recordings to measure cardiovascular system and skin conductance changes) and the somatic nervous system (motor expression in face, voice, and body). Therefore, we collect a new corpus, Rovereto Emotion and Cooperation Corpus (RECC), a task oriented corpus with psychophysiological data

registered and aligned with audiovisual data. In our opinion this corpus will allow to clearly identify emotions and, as a result, having a clearer idea of facial expression of emotions in dialogue. RECC is created to shade light on the relationship between cooperation and emotions in dialogues. To our knowledge this is the first dialogue corpus having audiovisual and psychophysiological data recorded and aligned together.

2. RECC Design

ECC is an audiovisual and psychophysiological corpus of dialogues elicited with a modified Map Task. The Map Task is a cooperative task used for the first time at the University of Glasgow and the HCRC group at Edinburg University (Anderson et al., 1991). The HCRC Map Task dialogues involved two participants. The two speakers sit opposite one another and each has a map which the other one cannot see. One speaker – designated as the *Instruction Giver* - has a route marked on his/her map while the other speaker - the *Instruction Follower* - has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their first session. All maps consist of landmarks – also called *features* - portrayed as simple drawings and labeled with a name. Our Map Task has some similarities with respect to the HCRC one. In front of them, the participants had both a map with a group of features. A number of them are in the same position and with the same name, but the majority of them is in different positions or has names that sound similar to each other (e. g. Maso Michelini vs. Maso Nichelini, Fig. 1). The Giver must drive the other participant (the Follower) from a starting point (the bus station) to the finish (the Castle of Rovereto). Giver and Follower were both native Italian speakers and they did not know each other before the task. As in HCRC Map Task, our corpus interactions have two conditions: screen and no screen. In the screen condition a barrier was present between the two speakers. In the no screen condition a short barrier was placed between the speakers allowing Giver and Follower to see each others' face. Screen conditions were counterbalanced. The two conditions enabled us to test whether seeing the speakers face during interactions influences facial emotion display and cooperation (for the relationship between gaze/no gaze and facial displays see Kendon, 1967; Argyle & Cook, 1976; for the influence of gaze on coordination see Brennan, et al., 2008). Previous studies have shown that visual access to each others' non verbal behavior fosters a dyadic state of rapport that facilitates mutual cooperation (Argyle, 1990; Tickle-Degnen & Rosenthal, 1990). However, these findings do not establish whether facial cues actually are predictive of cooperation. A further condition, *emotion elicitation*, was added. In emotion elicitation conditions the Follower or the Giver can alternatively be a confederate, with the aim of getting the other participant angry[1].

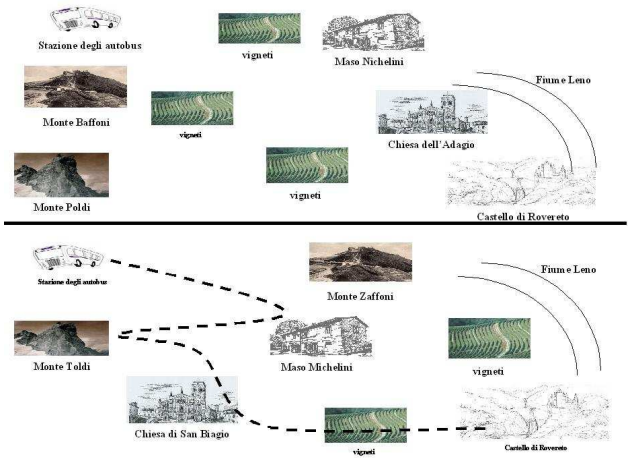

Figure 1: Giver and Follower Maps of RECC

3. Recording and Eliciting Procedure

RECC is made up of 20 interactions, 12 with a confederate, for a total of 240 minutes of audiovisual and psychophysiological recordings. During each dialogue, the psychophysiological state of non-confederate Giver or Follower is recorded and synchronized with video and audio recordings. The psychophysiological state of each participant has been recorded with a BIOPAC MP150 system. Audiovisual interactions were recorded with 2 Canon Digital Cameras and 2 free field Sennheiser half-cardioid microphones with permanently polarized condenser placed in front of each speaker.

The recording procedure of RECC was influenced by the work of Anderson, Linden and Habra (2005). They investigated the physiological arousal due to acute anger provocation. Before starting the task, we recorded the *baseline condition* of the participant. Specifically we recorded participants' psychophysiological outputs for 5 minutes without challenging them. Then the task started and we recorded the psychophysiological outputs during the interaction first three minutes that interaction occoured which we called the *task condition*. Soon, the confederate started challenging the other speaker with the aim of getting him/her angry.

Two groups of subjects were recorded. The first group consisted of 14 Italian native speakers (average age=28.6, dv=4.36) matched with a confederate partner. During these sessions, the confederate (the same person in all the interactions) performed uncooperative utterances in carefully controlled circumstances by acting negative emotion elicitation lines at minutes 4, 9 and 13 of the interaction.

The following lines were given by the confederate when acting the Follower role:

[1] All the participants had given informed consent and the experimental protocol was approved by the Human Research Ethics Committee of University of Trento.

- *"You are sending me in the wrong direction, try to be more accurate!";*
- *"It's still wrong, you are not doing your best, try harder! Again, from where you stopped";*
- *"You're obviously not good enough at giving instructions".*

A control group of 8 pairs of participants (average age=32.16, dv=2.9) were also recorded while playing the Map Task with the same maps. Eye contact, communicative role (Giver and Follower) and gender (male or female) conditions were counterbalanced.

Our hypothesis is that the confederate's uncooperative utterances would lead to a reduced level of cooperation in the other participant. To test it, we first need to check if the eliciting protocol adopted caused a change in participants' heart rate and skin conductance. In Fig. 2 we show the results of a 1x5 ANOVA executed in confederate condition. Heart rate (HR) is confronted over the five times of interest (baseline, task, after minute 4, after minute 9, after minutes 13), that is to say just after emotion elicitation with the script. A HR x Time ANOVA showed a significant effect of Time ($F(4, 8)=2.48$, $p<.001$), meaning that HR changed between task beginning and the three sentences in the script. In the control group session, in addition to a baseline measurement, HR was measured 3 times at equal intervals dung the interaction. A HRxTime ANOVA showed the effect of Time was non-significant ($F(3,5)=3,28$, $p<.117$). So, HR is significantly different in the confederate condition, meaning that is to say that the procedure to elicit emotions allows recognition of different psychophysiological states with respect to the non confederate condition. Moreover, the indicated HR values confirmed the ones found by Anderson and collegues (2005).

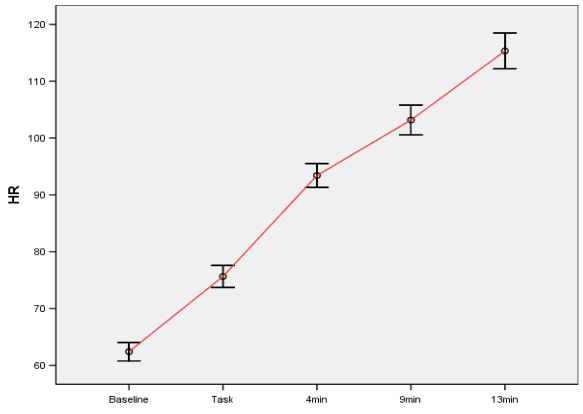

Figure 2: 1x5 ANOVA on heart rate (HR) over time in confederate condition in 12 participants

According to Tassinary and Cacioppo (2000), it is impossible to assess which emotion arises based on psychophysiological data alone. In fact, HR and skin conductance are signals of arousal. So, a high arousal can be due to emotions characterized by high arousal and high valence such as happiness or high arousal and low valence such as anger. Therefore, a 7 points Modified Self-Assessment Scale (adapted from Bradley & Lang, 1994) was completed by all participants. The aim was to obtain the emotive valence assessment measuring the polarity (positive to negative) of the emotion felt by each speaker toward his/her partner during the interaction. Subjective valence ratings were measured by having the 14 participants complete a 7.5 cm visual analogue emotion rating scale form.

From the inspection of skin conductance values (Fig. 3) there is a linear increase of the number of peaks of conductance over time. This can be due to two factors: emotion elicitation and also an increase in task difficulty leading to higher stress and therefore to an increasing number of skin conductance peaks.

The polarities of the rating scale were counterbalanced for left and right. 43% of the participants rated the experience as quite negative, 29% rated the experience as almost negative, 14% of participants rated it as negative and 14% as neutral. The 2 participants out of 14 reporting a neutral experience were discarded from further analysis.

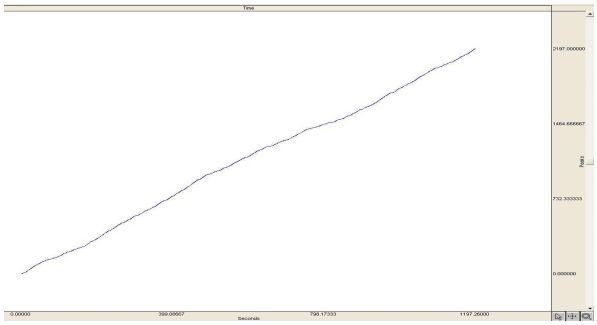

Figure 3: Number of skin conductance positive peaks over time in confederate condition in 12 participants

4. Corpus Transcription and Annotation

RECC corpus consists of manually produced orthographic transcriptions for each speaker and in addition, it is time aligned with all the communication modalities. Videos were imported using the ANViL (ANnotation of Video and Language) software (Kipp, 2001).

Orthographic Transcriptions of the interactions were done adopting a subset of the conventions applied to the transcription of LUNA project speech corpus (Rodriguez et al., 2007). All possible spontaneous speech effects were transcribed such as disfluencies, hesitations, repetitions, grammatical and pronunciation errors, and filled pauses. Two transcribers converted the recordings into plain texts. Every conversation was divided into turns related to the Giver and the Follower.

Six independent coders annotated 10 videos (6 belonging to the confederate sessions and 4 to the control session) following the guidelines reported in RECC Annotation Manual. Two coders repeated the annotation after one month to ensure its stability across time. All the annotators were Italian native speakers and only two of them had previous experience as coders. Cooperation and emotion were analyzed using a coding scheme based on

the decomposition of the several factors underlining an emotion. In particular, they did not refer to emotive terms directly. In fact every annotator had his/her own representation of a particular emotion, which could be different from the one of another coder. This representation can be a problem especially for the annotation of blended emotions, which are ambiguous and mixed by nature. In general, the analysis of non verbal features requires a different approach compared with other linguistic tasks. This is because multimodal communication has multiple semantic levels. For instance, a facial expression can deeply modify the sense of a sentence, such as in humor or irony. The reliability test we run had kappa scores between 0.79 and 0.80

5. Public releases

To the date, RECC is a free resource and can be asked by email to: federica.cavicchio@gmail.com. It is completely The psychopsysiological recordings are not publicly available, since the HR and the skin conductance data were collected and analyzed with Acknowledge®, a Biopac's licensed software. At the following link www.clic.cimec.unitn.it/RECC one can find reports of the documentations on the corpus collection methodology and the coding scheme.
The RECC annotation manual is available at http://www.clic.cimec.unitn.it/RECC, together with an XML file consisting of the ANViL specification file of the scheme.

6. Conclusion

RECC is a unique resource considering the way it was collected and the phenomena it challenged. It is the first multimodal corpus that includes audiovisual recordings aligned with psychophysiological data. RECC was built with the purpose to investigate linguistics, pragmatics and emotions in a dialogue setting. Our expectation is that researchers will obtain from the RECC elicitation method and the RECC annotation scheme a range of features that are necessary for the progress in the domain of multimodal dialogue studies. RECC coding scheme is another important step towards the creation of annotated multimodal resources which are crucial to investigate multimodal communication. Particularly, RECC coding scheme can aid exploring how different emotive sets (positive or negative) modify cooperation in different cultural settings; how turn management and sequencing strategies are expressed in different cultural settings; how gaze can enhance or disrupt cooperation; how emotions modifies the multimodal communicative channels.
Corpora annotated according to the RECC coding scheme represents useful resources to model back-channel, turn management and facial expressions of multimodal agents. Our findings will be hopefully taken into account in order to guide the design of Human Computer Interfaces.

7. References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P., (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena, Language Resources and Evaluation, 41, 273--287.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P., (2006). A Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. In Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., & Pianesi, F. (eds.) *Multimodal Corpora: From Multimodal Behavior Theories to Usable Models*, 38--42.

Anderson, J. C., Linden, W., & Habra, M. E., (2005). The importance of examining blood pressure reactivity and recovery in anger provocation research. International, *Journal of Psychophysiology*, 57,159--163.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., Weinert, R., (1991). The HCRC Map Task Corpus. *Language and Speech*, 34, 351--366

Argyle, M., (1990). The Biological Basis of Rapport. *Psychological Inquiry*, 1, 296 -- 300.

Argyle, M., & Cook, M., (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

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, 49--59.

Brennan, S.E., & Clark, H.H., (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 1482--1493.

Callejas, Z. & López-Cózar, R., (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, 50, 416--433.

Carletta, J., (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41, 181--190.

Cavicchio F., Poesio M. (2009). Multimodal Corpora Annotation: Validation Methods to Assess Coding Scheme Reliability, in M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen (Eds.), *Multimodal Corpora LNAI 5509*, Berlin: Springer-Verlag, 109--121.

Craggs, R., & Wood, M., (2004). A Categorical Annotation Scheme for Emotion in the Linguistic Content of Dialogue. In *Affective Dialogue Systems*, Elsevier. pp. 89--100.

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

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowi, R., Savvidou, S., Abrilian, S., & Cox, C., (2005).

Multimodal Databases of Everyday Emotion: Facing up to Complexity. In *9th European Conference on Speech Communication and Technology* (Interspeech'2005) Lisbon, Portugal, pp. 813--816.

Kendon, A., (1967). Some Functions of Gaze Directions in Social Interaction. *Acta Psychologica*, 26, 1--47.

Kipp, M., (2001). ANVIL - A Generic Annotation Tool for Multimodal Dialogue. In Eurospeech 2001 Scandinavia *7th European Conference on Speech Communication and Technology*.

Kipp, M., Neff, M., & Albrecht, I., (2006). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. In Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., & Pianesi, F. (eds.) In *Multimodal Corpora: From Multimodal Behavior Theories to Usable Models*. (pp. 24-28). Berlin: Springer Verlag

Magno Caldognetto, E., Poggi, I., Cosi, P., Cavicchio, F., & Merola, G., (2004). Multimodal Score: an Anvil Based Annotation Scheme for Multimodal Audio-Video Analysis. In Martin, J.-C., Os, E.D., Kühnlein, P., Boves, L., Paggio, P., & Catizone, R. (eds.) *Proceedings of Workshop Multimodal Corpora: Models of Human Behavior for the Specification and Evaluation of Multimodal Input and Output Interfaces*, pp. 29--33.

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K. & Abrilian, S., (2006). Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviors: Validating the Annotation of TV Interviews. In *Fifth international conference on Language Resources and Evaluation* (LREC 2006), Genoa, Italy.

Pianesi, F., Leonardi, C., & Zancanaro, M., (2006). Multimodal Annotated Corpora of Consensus Decision Making Meetings. In Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., & Pianesi, F. (eds.) *Multimodal Corpora: From Multimodal Behavior Theories to Usable Models*, pp. 6--19.

Poggi, I., & Vincze, L., (2008). The Persuasive Impact of Gesture and Gaze. In Martin, J.-C., Patrizia, P., Kipp, M., & Heylen, D., (eds.*) Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*,. Berlin: Springer Verlag pp. 46--51.

Reidsma, D., & Carletta, J., (2008). Reliability Measurement without Limits. *Computational Linguistics*, 34, 319--326.

Rodríguez, K., Stefan, K. J., Dipper, S., Götze, M., Poesio, M., Riccardi, G., Raymond, C., & Wisniewska, J., (2007). Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus. In *Proceedings of the Linguistic Annotation Workshop at the ACL'07* (LAW-07), Prague, Czech Republic

Tassinary, L. G., & Cacioppo, J. T., (2000). The skeletomotor system: Surface electromyography. In Tassinary, L.G., Berntson, G.G., Cacioppo, J.T. (eds) *Handbook of psychophysiology.* New York: Cambridge University Press pp. 263--299.

Tickle-Degnen, L., & Rosenthal, R., (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1, 285--293.

The emotional and communicative significance of head nods and shakes in a naturalistic database

R. Cowie [1], H. Gunes[2], G. McKeown[1], L. Vaclavu-Schneider[1], J. Armstrong[1], and E. Douglas-Cowie [1]

[1.] Queen's University, Belfast

[2.] Imperial College, London

E-mail: r.cowie@qub.ac.uk, g.mckeown@qub.ac.uk, e.douglas-cowie@qub.ac.uk

Abstract

Head nods and shakes have been extracted from the SAL audiovisual database of spontaneous emotionally coloured dialogue. The dataset contains 154 nods and 104 shakes. Two trained observers rated them on multiple dimensions derived from linguistics on one hand, and the psychology of emotion on the other. One used audiovisual presentation, the other visual only. There was agreement on affective, but not linguistic significance – suggesting that the latter depends on speech context rather than the manner of movement per se. A few seem to form discrete types, but for the most part classical dimensional models of emotion captured the affective variation well.

1. Introduction

The phrase 'emotion in action and interaction' was a repeated refrain in the HUMAINE project. Research is gradually coming to grips with the specifics of the way emotion enters into interactions. This paper is based on work being done in the successor project SEMAINE, whose aim is to create agents capable of engaging in sustained emotionally coloured interactions.

One of the effects of work in that area is to challenge standard divisions. This paper focuses on an area where a particularly complex set of divisions comes into play. The general domain that it deals with is backchannelling, which is generally thought of as a linguistic function. But whereas spoken language is usually thought of as a primarily acoustic phenomenon, a large part of backchannelling is visual. The paper considers the most obvious visible components of backchannelling, that is, head nods and shakes. Although the framework in which these gestures are usually analysed is (in a broad sense) linguistic, subjectively, at least part of their significance would seem to be emotional. Dictionaries typically give one of the meanings of 'shake' along the lines of 'To brandish or wave, especially in anger', and it is hard to believe that there is no relationship between that and shaking the head to signify a negative reaction. The upshot is that the area brings together multiple modalities and multiple types of significance in an intriguing package.

Although the issues are complex, the end product is straightforward: a database of over 250 head movements extracted from spontaneous interactions, with associated labels that capture those attributes of each movement that seem to be most salient to human observers. It provides a basis for research on either recognition or synthesis of appropriate types of head movement during interaction.

2. Analyses of nods and shakes

Head movements have been viewed in various ways among the computational community and related disciplines. Historically, the usual practice been to treat head movements during conversation as noise to be ignored as best one can. Until recently, speaking avatars did not generally move their heads; and if people speaking to them made head movements (which was discouraged), the main response was to look for ways of recovering facial expression in spite of the complication produced by head movement.

An alternative which has become widely recognised is to regard head movement during speech as a default whose presence is not particularly informative, but whose absence is. The background to that position was provided by the motor theories of investigators such as Hadar (1984a,b) who argued that large scale movements during speech create a favourable environment for the subtle, co-ordinated actions required to produce speech as such. The idea was highlighted by evidence that suppression of default accompanying movements was associated with deliberate (and deceptive) manipulation of the communication process (Cohn et al 2004).

A second major alternative is linguistic. Head movements have been regarded as an integral part of the concept of backchannelling since the concept emerged (Yngve 1970, Duncan 1972). Nods in particular were seen as an integral part of the mechanism by which speakers manage control and exchange of the 'floor'. That approach was elaborated in an influential paper by McClave (2000), who distinguished nine types of linguistic function for nods. These are described in the method section.

That conception has influenced computational research in general, and the SEMAINE project in particular (Heylen et al, 2007). SEMAINE aims to synthesise agents that can hold a sustained, emotionally coloured conversation with a user. One of the key ways in which it creates a sense of interaction is by having the agents make head movements, and respond to the user's head movements. The database described here was created to support the development of that aspect of SEMAINE.

Although the linguistic perspective influenced SEMAINE directly, it is clearly not the only possible option. Two others will be mentioned here.

Psychology has a long-standing interest in interpersonal behaviours whose function seems to be to show 'convergence' between the interactants. There are famous examples involving posture (Beattie & Beattie 1981), but there has been a recent surge of interest in behaviours that show temporal alignment (often described as entrainment) (Varni et al 2009). That approach invites the idea that head movements support synchronisation, serving both to achieve and to display temporal coherence.

It also is natural to assume that head movements may express affective content. There is an obvious connection with approaches that describe affect in terms of two dimensions, valence and arousal. To a first approximation, it would seem likely that a nod expresses positive affect towards the other party, whereas a shake expresses negative affect; and there is a relationship between the energy of the gesture and the arousal level.

Last, but not least, it should be noted that cultural factors are a major unknown. It is certainly true that some head movements take on some specific meanings in certain cultures (Brodsky & Griffin 2009). What is not clear is how deep and wide the influence of cultural factors is. This study does not try to answer that question, though the techniques that it describes might be relevant to doing so.

3. The study

The material to be considered was extracted from recordings of interactions using the SAL paradigm (Douglas-Cowie et al 2008). SAL is short for 'sensitive artificial listener'. The technique generates emotionally coloured conversation between a user and "characters" whose responses are stock phrases keyed to the user's emotional state rather than the content of what he/she says. The model is a style of interaction observed in chat shows and parties, which aroused interest because it seemed possible that a machine with some basic emotion-detection capabilities could achieve it. In earlier versions of SAL, designed to provide training data, a person emulated the SAL characters. There are now versions where the characters' speech and visible gestures are generated automatically. Recordings are available via http://www.semaine-db.eu/. SAL

Nods and shakes were extracted from interactions between a user and the person emulating the characters. The core task was to provide a description of each item that captured its functional significance as perceived by humans. In order to do that, it was necessary to address conceptual questions about the kind of framework that best captures the meaning that people attach to these movements. Two main levels of question are addressed. The first is whether distinctions between head movements are best captured in terms of linguistic categories (using McClave's system as the best developed) or affective descriptions. The second is whether the distinctions are best expressed in terms of categories (i.e. nods fall into n types) or dimensions (i.e. they lie at different points on n continua).

3.1 Method

3.1.1 Rating scales

The rating scales involved two parts. The first part used selected components of the system that has been developed for SEMAINE. It covers a range of descriptors, from the classical dimensions used to describe pure affect (arousal and valence) to terms that are purely cognitive (understanding and agreement). Between these are terms with both social and affective implications – 'solidarity' and 'antagonism', drawn from the categories developed by Bales (2000), which relate mainly to the valence dimension; and 'at ease', which relates mainly to arousal.

The second part used the linguistic categories proposed by McClave, i.e. inclusivity; intensification; uncertainty; direct quotes; expression of mental images of character; deixis and referential use of space; lists or alternatives; lexical repairs; backchanneling requests.

3.1.2 Rating procedure

All items were rated by two observers. They were students working on a year-long project, who were given prior training in the meanings of the categories as a preparation for the exercise.

Since there is a question over the role of linguistic information, we adopted the simple solution: one rater used audiovisual presentation, the other used visual only. Order of presentation was randomised, using different orders for the two raters.

The SEMAINE-derived components were used for both nods and shakes, the McClave components for shakes only.

4. Analysis

Ratings for nods and shakes were analysed separately, The same basic issues were covered in both. For each category, agreement between the two raters was examined. Note that since one rater had linguistic information and the other did not, what agreement indicates is that ability to assign that category does not depend radically on the presence of linguistic information. The two ratings were then averaged, and a second level of analysis was applied to the resulting averages. The first step was cluster analysis. Two step cluster analysis was used, as the most straightforward option. The results of that analysis were then studied graphically to establish whether some clusters might be better regarded as portions of a continuum (typically the upper and lower extremes). Where there seemed to be evidence of a continuum, factor analysis was used to gauge the number of dimensions present and the proportion of the variance that they accounted for.

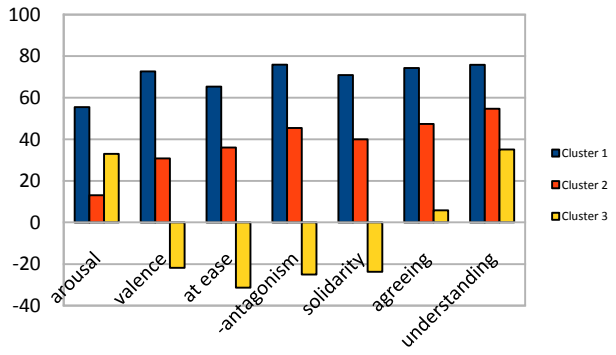

Figure 1: coordinates of cluster centroids for nods

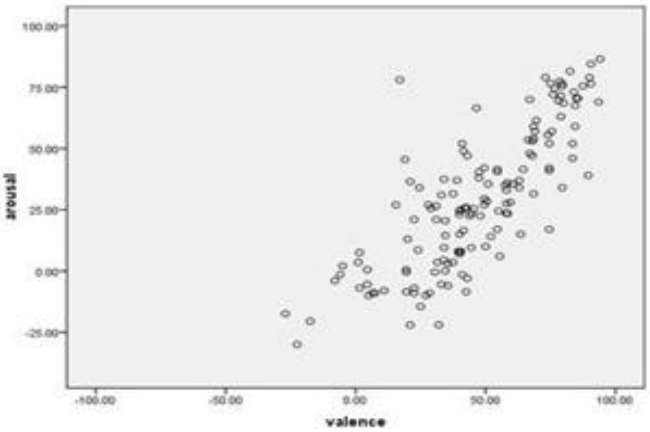

Figure 2: valence and arousal in nod clusters 2 & 3

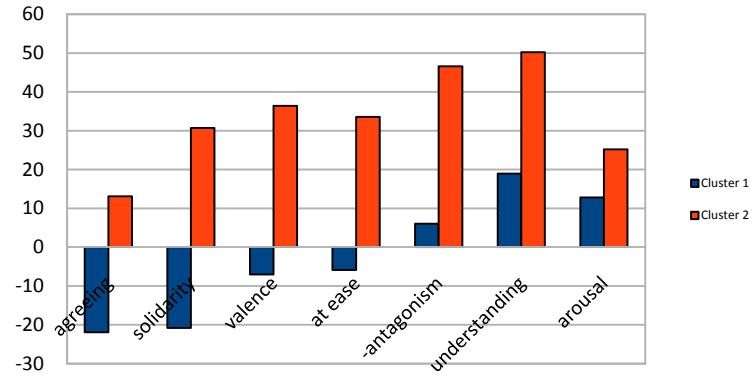

Figure 3: coordinates of cluster centroids for shakes

Variable	factor 1	factor 2
valence	0.86	0.24
arousal	0.31	0.91
agreeing	0.68	-0.34
at ease	0.83	0.1
solidarity	0.87	-0.27
antagonism	-0.8	0.2
understanding	0.8	0.08

Table 1 loadings of shake factors on the items

4.1 Nods

Inter-rater agreement was significant for most of the SEMAINE variables, though the strength of the relationship varied. It was highest for arousal (r =0.585), and valence (r=0.393), and low (but still significant) for solidarity (r=0.232), antagonism (r=0.190) and agreement (r= 0.213). Agreement was non-significant for understanding and at ease. The natural interpretation is that seeing a nod provides good information about affect, but relatively little about the more interpersonal and cognitive issues.Cluster analysis identifies three clusters. Figure 1 shows the coordinates of the cluster centroids. The profiles of clusters 1 and 2 are almost parallel, which is what would be expected if the clusters actually represented upper and lower ends of a single continuum. Cluster 3 is very different. It is marked by arousal and understanding, along with a range of negative evaluations – in other words, the nods convey that a message is understood and rejected.

To clarify the meaning of clusters 1 and 2, the 'understand and reject' nods were removed and factor analysis was applied to the remaining points. It recovered one factor, which corresponds to a well-established concept in emotion research, 'positive activation' (Watson & Tellegen 1985): that is to say, there is a continuum from low activation and neutral to high activation and positive. Figure 2 shows the distribution with respect to the two affective variables. It seems much more natural to regard these nods as a single continuum than as two clusters.

4.2 Shakes

The pattern of inter-rater agreement for shakes was broadly similar to the pattern for nods, with clearly significant agreement for arousal (r=0.605), intermediate for valence (r=0.264), solidarity (r=0.324), and antagonism (r=0.307); and marginal or non-significant relationships for agreement, understanding, and at ease.

Again, the natural reading is that what the appearance of shakes signals is mainly affective. An interesting additional point can be made because raters were asked to give not only their rating, but also their confidence in it. The items of which the rater with audiovisual information was most confident were 'understanding' and 'at ease', suggesting that the reason for inter-rater differences on these items is not that the information is poor, but that the audio provides very good information for them.

Cluster analysis identifies two clusters. Figure 3 shows the coordinates of the cluster centroids. However, if we ignore arousal, again, the profiles are almost parallel, suggesting that they may represent upper and lower ends of a single continuum. If so, arousal follows a different pattern. Factor analysis confirms that reading. It finds two dimensions. The loadings, shown in Table 1, show that they correspond admirably to the classical affect dimensions of valence and arousal. That suggests that the information in shakes is even more straightforwardly affective than that information in nods.

Analysis of the linguistically motivated categories reinforces that point. Most of the categories were not applied with any consistency at all. Specifically, for seven of the nine categories, the number of clips where both raters agreed that the category applied was two or less. The two exceptions were intensification and uncertainty, which it is reasonable to regard as the most affective of the categories. The lack of agreement on the others, which are more straightforwardly linguistic, has a straightforward interpretation: it seems very likely to mean that what marks these functions is not the appearance of the shake per se, but its relationship to speech.

5. Discussion

The concrete outcome of the research is a database containing substantial numbers of nods and shakes from spontaneous, emotionally coloured interactions, and a variety of labellings. That provides a resource for research interested in either learning or synthesising head movements during interaction.

The conceptual outcome can be expressed in terms of the way the various labellings included in the database can be understood. There is a strong tendency to assume that gestures like nods and shakes should be understood in terms of categories. Membership of clusters is given, and it does seem to be a useful descriptor for one, rather small group of nods, those that convey “message understood and rejected”. However, in most cases, the natural descriptor follows one of the classical patterns described by the psychology of emotion – positive activation for nods, and valence/activation space for shakes.

Describing these patterns highlights an issue which, to the best of our knowledge, has not been brought into focus before. It is the role of statistical reduction techniques in labelling. It is a standard procedure in psychology to translate responses on a number of raw rating scales into a smaller number of scores on (ex hypothesi) more basic dimensions. Derived measures of that kind are generated by the factor analyses described here, and the results are included in the databases. It would be consistent with practice elsewhere in psychology to think of that kind of variable as a more natural source of information than responses to individual items.

Linked to doubts about classification are doubts about the way paradigms from linguistics apply to non-verbal communication. The database includes descriptions using McClave's categories. It is not in dispute that the categories are useful. However, there clearly is reason to question the basis on which they are assigned. It has not been emphasised, but it is possible, that the categories simply cannot be assigned with much consistency. However, it seems more likely that the reason for the inter-rater differences is that assigning these categories is not a matter of classifying the head movement as such, but of gauging its relationship to what is being said.

A second level of relationality has not been addressed directly, but the results raise questions about it. There are reasons to predict that the meaning of a head movement will only be apparent in the context of the other party's movements. That appears to be at most part of the picture. It may be that relative timing can change the perception of a head movement, but there seem to be conclusions that people can draw with high confidence without considering that context.

Acknowledgements

Preparation of this paper was supported by the FP7 projects SEMAINE and SSPnet.

References

Bales, R.F. (2000) Social Interaction Systems: Theory and Measurement. New Brunswick, NJ: Transaction.

Beattie, G. and Beattie, C. (1981) Postural congruence in a naturalistic setting Semiotica 35 (1-2), 41–56

Brodsky, S.L & Griffin, M.P. (2009) When jurors nod The Jury Expert 31(6) 38-40

Cohn, J.F. Reed, L.I. Ambadar, Z. Xiao, J. & Moriyama, T. (2004) Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior, Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '04), vol. 1, pp. 610-616.

Douglas-Cowie, E. Cowie, R. Cox, C. Amir, N. & Heylen, D. (2008) The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC ’08), pp. 1–4, Marrakech, Morocco, May 2008.

Duncan, S. (1972) Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23: 283-292.

Hadar, U., Steiner, T.J. Grant E.C. and Rose, F. C.

(1984a). The timing of shifts of head postures during conversation. Human Movement Science 3: 237-245.

Hadar, U., Steiner, T.J. and Rose, F. C. (1984b) The relationship between head movements and speech dysfluencies. Language and Speech 27: 333-342.

Heylen, D., Nijholt, A., & Poel, M. (2007) Generating Nonverbal Signals for a Sensitive Artificial Listener. Verbal and Nonverbal Communication Behaviours, pp. 264-274.

McClave, E.Z. (2000) Linguistic functions of head movements in the context of speech Journal of Pragmatics 32 (7), 855-878

Varni, G., Camurri, A., Coletta, P. Volpe, G. (2009) Toward a Real-Time Automated Measure of Empathy and Dominance. International Conference on Computational Science and Engineering, 2009 vol. 4, pp. 843-848,

Watson, D., Tellegen, A. (1985). Toward a consensual structure of mood. Psychological Bulletin, 98, 219–235

Yngve, V. 1970. On getting a word in edgewise. Papers from the 6th regional meeting of the Chicago Linguistic Society, 567-578.

.

Annotation of Affective Interaction in Real-life Dialogs Collected in a Call-center

Christophe Vaudable, Nicolas Rollet, Laurence Devillers
LIMSI-CNRS
Orsay, France
E-mail: christophe.vaudable@limsi.fr, devil@limsi.fr

Abstract

The originality of this paper is to propose an annotation scheme including new affective dimensions linked to social and affective interactions. These annotations allow us to describe more precisely the conversation between two speakers considering the theories of conversational analyses. The work is done in the context of the French national VoxFactory project.

1. Introduction

In the field of affective computing one of the recurrent problems is the lack of experimentation on "real-life" data. Acted corpora are often composed of samples of very prototypical emotions and are not representative of reality.

In this study we have focused our attention on data collected by a call center. This kind of corpus is very rich and presents many challenges. In the past we have already worked on several corpora collected in such call centers [1] (medical, stock exchange, etc.). The work presented here was done in the context of the French national VoxFactory project[1]. The corpus we have used for this study comes from a previous project: the CallSurf project [2]. This new study concerns the customer services of French companies. Talk during social interactions naturally involves the exchange of propositional content but it also involves the expression of interpersonal relationships, as well as displays of emotion, affect, interest, etc. In the specific case of interaction in call centers, the agents deploy several different strategies to satisfy their customers by making commercial propositions and offering solutions.

Our research focuses on the analyses of emotional behavior which occurred during interactions over the phone between an agent and a customer. Many annotation schemes about emotions focus only on the annotation of one speaker turn without taking into account all interlocutors at the same time; so conversations are not analyzed as interactions between several people but as a series of speaker turns, analyzed separately. This paper aims to propose a new annotation scheme including affective dimensions linked to the social and affective interactions. Our annotation scheme is based on previous schemes [3] [4] including labels and emotional dimensions such as Pleasure, Arousal and Dominance, to which we have added new dimensions to describe affective involvement : reaction, induction and the possibility to link segments between them.

Due to plurality of the emotional behavior in real-life interactions, we inititially selected only a part of the 150 hours corpus collected in the call center (section 2). Then we have made a first annotation experiment to test and improve our annotation scheme on a subset of this corpus. Section 3 explains our annotation scheme. In this experiment we also compared two different ways to annotate the data (continuous and discrete, section 4). Initial results are finally given in section 5 on the correlation between dimensions and labels.

2. Corpus

The data was collected from one of the EDF (French electricity utility) call center. The CallSurf Corpus [2] was composed of about 1,000 hours of dialogs composed of more than 10,000 calls. The sparseness of the emotional content in real-life data led us to select the more emotional parts of the collection. From 150 hours of transcribed data, we selected 15 hours which corresponds to 77 phone calls. The duration of the calls is between 1 and 30 minutes. To select these 77 calls we have used a subset of corpus provided by TEMIS and EDF (partners of the VoxFactory project). This subset of the global corpus is selected by an automatic lexical analysis which indicates dialogs containing emotionally marked words. We then manually selected the 77 calls from this subset from an acoustic analysis. This step will be done automatically in the next step of the project.

The EmoVox corpus is now composed of 243 extracts from those 77 calls. Each extract represents the more emotionally part of each call. EmoVox is composed of 2 h 42 min of speech: the 243 extracts have been segmented in 2,990 segments. We have extracted 5 calls from this sub corpus and called it EmoVox1. The 5 extracted dialogs have been segmented into 36 extracts. These 36 extracts have been divided into 450 segments. Each emotional segment is, for a given speaker, a homogeneous emotional turn (which represents one or less than one speaker turn). We have excluded overlapped segments (two people speaking in the same time); the proportion of this kind of segments in EmoVox1 is about 20% (92 segments).

[1] Cap digital French national project founded by FUI6

All the extracts of those dialogs are very emotional with negative and positive emotional behaviors and allows us to test our annotation scheme and build an annotation guide.

3. Annotation

3.1 Annotation Scheme

Real-life emotions are more complex to annotate than basic acted emotions. Our multi-level scheme allows us to describe emotions and new affective dimensions linked to the interaction. The main difficulty of this representation is to find the useful levels of description in terms of granularity and temporality.

We have included both verbal categories and abstract dimensions to describe emotion in order to study their redundancy and complementarities as in [3-4].

3.2 Affective dimension in interaction

Data are collected from phone conversations, so we have to take into consideration the fact that this is a social interaction; preliminary research on conversational analysis has shown:
(1) The participants try to organize their speaker turns using prospective and retrospective verbal practices in order to give their interlocutor a coherent sequence of speaker turns that can be recognized as a conversation [5];
(2) Affective expressions are linguistic resources that emerge from a conversation. They are used as a sign of commitment to the interlocutor.

These two points were used as a theoretical basis for the construction of two new dimensions of affective implication and an annotation tool that we will present in the following paragraphs.

The development of our "EmoTool" has therefore taken into account some specificities of a conversational analysis by allowing to link the different speaking turns of a conversation between them.

The dimensions that describe the affective implication can be defined as follows:

- Reaction / induction axes: an emotional segment is described according to its propensity of being prospective or retrospective, i.e. its degree of connection to the precedent or following segment.
- Intersegment relation: Considering statements (1) and (2) we added to EmoTool the possibility of linking the different segments of an interaction between them to enable the description of coordinated verbal actions as described in [5-6].

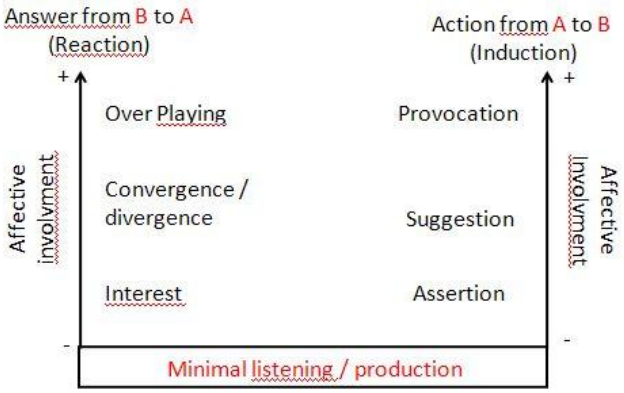

Figure 1: Dimensions of affective implication

For instance we can see that the lowest levels of both axes of affective implication are linked with a minimal listening (or production) of a speaker in a conversation. A high level of induction on the right axis is correlated with a kind of provocation (threat or seduction for instance). The important thing to remember concerning this axis is that the aim of the induction is to create a reaction during the next turn of the conversation.
A high degree on the reaction axis describes for instance behavior like an emotional overbid. Contrary to the induction, the reaction represents the impact of a previous speaker turn on a current turn. In the following we present illustrative examples of (translated) speaker turns:

Agent: *"If you don't pay the bill before tomorrow we will be obliged to turn off you electric installation"* (High degree of induction) (Pleasure/Arousal/Dominance Values: -1/0/0 – Induction: 4 Reaction: 3)
Client: *"Don't threaten me!"* (High degree of reaction) (PAD Values: -2/1/-1 Induction: 4, Reaction: 4)

These new dimensions of annotation were tested by 3 coders in an annotation experiment of the sub-corpus EmoVox1.

3.3 EmoTool

For the needs of the study we next introduce our annotation tool. Existing tools like [7], [8] or [9] do not provide one package all the functionality needed for our purpose. Our software allows us to annotate PAD dimensions [10] and the dimension of "affective implication". Both of these dimensions can be linked with labels (see figure 2 for the complete list of labels used).

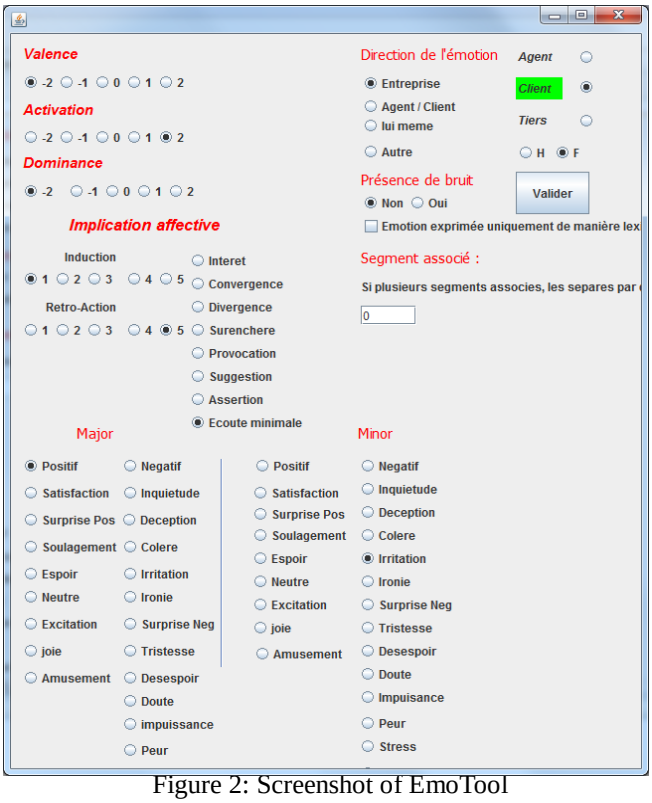

Figure 2: Screenshot of EmoTool

EmoTool allows a continuous annotation in time: the annotator has to move her or his mouse on the screen and all the movements of the mouse are recorded in a file with the corresponding time code of the audio clip.

4. Inter-agreement Annotation

4.1 Continue vs. discrete dimension annotations

The EmoVox-1 corpus was annotated continuously and discretely by 3 annotators for the continuous method and 4 annotators for the discrete method (2 female annotators aged 27 and 32 years and 2 male, aged 26 and 33 years). The rates of agreement of each annotation and the time were compared with the aim to find the most efficient way to annotate the overall corpus.In terms of time spent to annotate the corpus, the continuous and discrete methods are very similar. A difference of only few minutes, over about 8 hours of annotation work, was found. The rate of agreement for the three PAD dimensions appears in the tables 1 and 2 using the continuous method (table 1) and the discrete one (table 2). To compute the rate of agreement for the continuous annotation we cut the space into three equal zones for each dimension (positive / negative / neutral), and observed in which part of the emotional space the mouse cursor was at a given time corresponding to the segments.

Dimension	M1-M2	M1-W1	M2-W1	**Av.**
Pleasure	0.79	0.80	0.69	**0.76**
Arousal	0.58	0.62	0.74	**0.65**
Dominance	0.80	0.66	0.56	**0.67**

Table 1: Rates of agreement for a continuous PAD annotation (3 coders: M1, M2: men, W1: woman)

Dimension	M1-M2	M1-W1	M2-W1
Pleasure	0.49/0.60/**0.71**	0.35/0.53/**0.71**	0.41/0.58/**0.73**
Arousal	0.42/0.49/**0.59**	0.32/0.38/**0.48**	0.38/0.45/**0.56**
Dominance	0.59/0.64/**0.71**	0.20/0.29/**0.41**	0.11/0.21/**0.34**

Dimension	M1-W2	M2-W2	W1-W2
Pleasure	0.21/0.45/**0.64**	0.22/0.46/**0.64**	0.39/0.64/**0.80**
Arousal	0.17/0.21/**0.26**	0.16/0.23/**0.30**	0.36/0.46/**0.59**
Dominance	0.22/0.31/**0.41**	0.22/0.3/**0.40**	0.08/0.22/0.**38**

Table 2: Kappa square coefficient for discrete PAD annotation (4 coders: M1, M2: men, W1, W2: women)

Our dimensions allow us to provide κ by Cohen's unweighted, linear and quadratic weighting. A monotonic increase going from unweighted to quadratic thereby indicates label confusions preferably in neighboring classes. Our results in table 1 and 2 show that the discrete and continuous methods can be quite similar for example for Pleasure. The discrete method shows more precise results and has been kept for the next part of the study because it is more adequate for the annotation of the affective involvement.

4.2 Labels' annotations

During the step of discrete annotation, judges had the possibility of choosing one or two labels (major-minor) corresponding to the emotion or mixtures of emotions they perceived in the listening phase. In this experiment, we have considered the first emotion annotated by the 4 coders. If we consider the annotation of more than one label, the agreement can be higher. For instance, anger/sadness and sadness/anger could be considered as the same coding which is not the case in our first computation of the agreement. In section 4.3, we will discuss the complex emotions presented in this corpus. We can see in table 3 the score of agreements for 5 Macro-classes (collection of emotions): the positive / negative and neutral classes and the scores for two classes (negative against the remaining classes). These scores are around 0.5 in average for macro-classes and 0.8 for valence which are quite high. The macro-classes are positive, fear, neutral, anger, sadness.

Labels	M1-M2	M1-W2	M2-W1	M1-W2	M2-W2	W1-W2	**Av.**
Macro-Classes	0.57	0.49	0.62	0.50	0.53	0.50	**0.54**
Pos/Neg/ Neut	0.73	0.66	0.76	0.64	0.68	0.75	**0.70**
Pos-Neut /Neg	0.82	0.79	0.86	0.82	0.81	0.82	**0.82**

Table 3: Rates of agreement for the labels' annotation.

4.3 Complex emotions

One of the specific characteristic of real-life data is the frequent presence of blended emotions. We can see in table 4 the percentage of blended emotions in EmoVox1 according to the judges:

	H2	W1	W2
% of blended emotion	49	62	69

Table 4: Percentage of blended emotions for each annotator

These scores represent the rate of blended emotions for all the segments of the corpus. On average the score is higher when we focus our attention on the client only (71% of blended emotions). Most of them are mixtures of positive or negative. We have also found some complex emotion composed by a negative and positive emotion. Table 5 depicts some examples of these complex emotions.

First emotion (Major)	**Second emotion (minor)**
disappointment	hope
Positive	irony
doubt	positive

Table 5: Examples of complex emotions

4.4 Affective involvement

This concept is composed by the 3 elements: Induction and Reaction and intersegment relation as described in section 3.2.

As first step we computed the score of agreement between 3 annotators on the dimensions of induction and reaction using only two possibilities: high (for value < 3) and low for value >= 3) as we can see in table 6

Dimension	M2-W1	M2-W2	W1-W2	**Avg.**
Induction	0.80	0.76	0.75	**0.77**
Reaction	0.70	0.77	0.73	**0.73**

Table 6: rates of agreement for affective involvement

The scores of agreement between the 3 annotators were quite good; we then focused our attention on the impact of the production of emotion of the agent (EDF employee) on the client and vice versa.

5. Affective divergence/convergence in the dialogs

It is meaningful to follow the divergence or convergence of the dialogs using the expression of satisfaction and dissatisfaction - this is one of our goals in the project VoxFactory. As a first step, we have analyzed the annotation of our corpus, in order to correlate the measures of new dimensions proposed with the more "classical" annotations with labels and PAD dimensions.

As we pre-selected data paying attention to not include too many neutral instances, the induction score is higher than 3 in EmoVox1. In this "mini corpus" the average of neutral segments is about 20.0%. In 87.5% of the cases the high value of affective involvement is produced by the client.

We observed the impact of a high value of induction (> 3) (cf. table 7) and reaction (cf. table 8) on the values of the PAD scale.

	Pleasure	Arousal	Dominance
Pos	1.56	1.44	1.00
Neg	-1.60	0.00	-1.30

Table 7: Impact of a high value of induction on the emotion production

	Pleasure	Arousal	Dominance
Pos	1.30	1.40	0.00
Neg	-1.65	-1.00	-1.50

Table 8: Impact of a high degree of reaction on the emotion production

As we can see, the values of the PAD scale are considerably high in the case of high reaction or induction. According to the fact that the affective involvement is an emotional response to the previous or next speaker turn, these results seems reasonable in particular for the client.

We have next computed the values of the two dimensions of affective involvement considering selected emotional labels (Neutral and Anger). One can observe that the scores of induction and reaction are strongly linked with those of valence, arousal and dominance. Table 9 provides illustrative examples of the values of induction and reaction for the emotions annotated in the corpus.

Macro Class	**P / A / D**	**Induction / Reaction**
Neutral	0.0 / 0.0 / 0.1	1.8 / 1.9
Anger	-1.4 / 1.4 / -1.2	3.4 / 3.5

Table 9: Impact of high values of affective involvement on PAD dimensions and selected emotional labels

We further investigated the score of induction and reaction for agents and clients. We found that in more than 90% of the cases the high value of induction (more than 3) are produced by the client. The observation is also valid for the reaction dimension with more than 85% of high values attributed to the client. These results seem relevant considering the context of a call center in which agents have to avoid as much as possible negative behaviors for obvious commercial reasons.

Considering our two dimensions of affective involvement we have focused our attention on the relation between the different segments of an extract and the impact of the affective involvement on the evolution of speakers' emotion production.

First we have computed the average distance between two segments in the corpus. We observed that a very large majority of these links have a distance of 1 only. On average a distance between 2 segments in EmoVox 1 is1.1, and 90% of the segments of the corpus are linked to at least another one due to the fact that we have selected emotional extracts of the dialogs. In EmoVox 1 9.6% of the segments are linked with several other segments (two or more) by at least one coder. It is often the case when a problem cannot be solved by the agent.

In a second step we have observed in a more precise way the situations in which the emotional involvement was mainly implemented. We have noticed that for the majority of the cases the emotional implication played a role in two distinct conversational processes:

1) Emotional alignment of one of the speakers on the other. This situation occurs when one of the speakers (mainly the customer) produces a very intense emotion during several speaker turns. Table 10 provides an example of emotional alignment on 3consecutive speaker turns:

Turn	Spkr	PAD Value	Ind	Reac	Annotations by the 3 coders (Major/Minor)
I	Agent	0/0/0	2	2	Neut/ Neg Neut/ Worry Neut/Neut
I +1	Client	-2/2/-2	4	3	NegSur/ Ang Ang/ Worry Ang/Neut
I +2	Agent	-1/1/-1	1	3	Neg/ Wor Neg/Neur Irr/Neut

Tables 10: Example of emotional alignment – Ind : Induction, Reac : Reaction
Neut: neutral, Neg-Sur: Negative surprise, Ang: anger, Irr: irritation, Wor : worry

As table 10 shows, the value on the PAD scale of the client (turn I+1) was rather extreme and linked with a very high degree of induction (all these values were high in consecutive past speaker turns). Progressively we can observe an increase in the reaction value of the agent. The result is an alignment of the emotional state of the agent on the customer's state.

2) The complementary case is the attempt of the agent to keep the control of the conversation. Contrary to the emotional alignment explained before, the agent shows a low score of reaction (generally lower than 2 on a scale of 5) as seen in table 11 :

Turn	Spkr	PAD Value	Ind	Reac	Annotations by the 3 coders (Major/Minor)
I	Agent	0/0/0	2	2	Neut/Neut Neut/Neut Neut/Neut
I +1	Client	-2/2/-2	4	2	Neg-Sur/ Neg Anger/Wor Anger/Neut
I +2	Agent	0/0/0	1	2	Neut/Neut Neut/Neut Neut/Neut

Tables 11: Example of the control maintaining strategy of the agent

In this case it can be observed that the value of affective involvement of the agent is very low in both dimensions. The agent keeps the situation under control with a very neutral attitude. Another example of these strategies for keeping the control on the conversation is found in table 12:

Turn	Spkr	PAD Value	Ind	Reac	Annotations by the 3 coders (Major/Minor)
I	Agent	0/0/1	2	2	Neut/Neut Neut/Neut Neut/Neut
I +1	Client	-2/2/-2	3	2	Ang/ Irrit Ang/ Wor Ang/Neut
I+2	Agent	1/0/1	2	3	Pos/Neut Neut/ Pos Neut/Neut

Table 12: another example of control's strategy

Here we can see that the agent tries to calm the customer with a positive attitude and speech (Example (translated): *"I'm here to help you sir."*).
These three examples are very prototypical and have been chosen to show the possibilities of affective involvement. All these results need to be confirmed by computing all respective values on the entire corpus, but provide an encouraging first step for the establishment of a new scale of emotion measurement in conversations.

6. Conclusions

In this paper we have introduced the new dimension of affective involvement. Its annotation allows us to describe more precisely the conversation between two speakers considering the theories of conversational analyses. The entire presented corpus will be annotated with this scheme in the near future. The addition of semantic and pragmatic information will be further necessary to fully exploit the potential of affective involvement.

7. Acknowledgements

The authors would like to thank EDF R&D for collecting the data, VECSYS for transcriptions of the dialogs, TEMIS and EDF R&D for corpus pre-selection, and the CAP DIGITAL competitiveness cluster.

8. References

1. Devillers, L. and L. Vidrascu, *Emotion detection in real-life spoken dialogs recorded in call center.* Journal of Neural Networks, numéro spécial 2005. **volume 18**.
2. Garnier-Rizet, M., et al., *CallSurf - Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content*, in *LREC*. 2008: marrakech.
3. Devillers, L., et al., *Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches*, in *LREC06*. 2006: GENOA
4. Devillers, L. and J.C. Martin, *Coding Emotional Events in Audiovisual Corpora*, in *LREC 2008*. 2008.
5. Sacks, H., *Lectures on conversation* 1992: Blackwell Publishing.
6. Sacks, H., E. Schegloff, and G. Jefferson, *A simplest systematics for the organization of turn-taking for conversation.* Language, 1974. **50**(4): p. 696-735.
7. Kipp, M., *Anvil - A Generic Annotation Tool for Multimodal Dialogue*, in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*. 2001: Scandinavia.
8. Cowie, R., et al., *Feeltrace: an instrument of recording perceived emotion in real-time*. 2000.
9. Barras, C., et al., *Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech*, in *LREC 98*. 1998: Granada, Spain. p. 1373-1376.
10. Merhabian, A., *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament.* Current Psychology: Developmental, Learning, Personality, Social, 1996. **14**: p. 261-292.

Presenting the VENEC Corpus: Development of a Cross-Cultural Corpus of Vocal Emotion Expressions and a Novel Method of Annotating Emotion Appraisals

Petri Laukka[1], Hillary Anger Elfenbein[2], Wanda Chui[3], Nutankumar S. Thingujam[4], Frederick K. Iraki[5], Thomas Rockstuhl[6], & Jean Althoff[7]

[1]Department of Psychology, Stockholm University, Stockholm, Sweden
[2]Olin Business School, Washington University in St. Louis, St. Louis, MO, USA
[3]Haas School of Business, University of California, Berkeley, CA, USA
[4]Department of Psychology, University of Pune, Pune, India
[5]United States International University, Nairobi, Kenya
[6]Nanyang Business School, Nanyang Technological University, Singapore
[7]UQ Business School, University of Queensland, Brisbane, Australia
E-mail: petri.laukka@psychology.su.se, helfenbein@wustl.edu

Abstract

We introduce the *Vocal Expressions of Nineteen Emotions across Cultures* (VENEC) corpus and present results from initial evaluation efforts using a novel method of annotating emotion appraisals. The VENEC corpus consists of 100 professional actors from 5 English speaking cultures (USA, India, Kenya, Singapore, and Australia) who vocally expressed 19 different affects/emotions (affection, amusement, anger, contempt, disgust, distress, fear, guilt, happiness, interest, lust, negative surprise, neutral, positive surprise, pride, relief, sadness, serenity, and shame), each with 3 levels of emotion intensity, by enacting finding themselves in various emotion-eliciting situations. In all, the corpus contains approximately 6500 stimuli offering great variety of expressive styles for each emotion category due to speaker, culture, and emotion intensity effects. All stimuli have further been acoustically analyzed regarding pitch, intensity, voice quality, and durational cues. In the appraisal rating study, listeners rated a selection of VENEC-stimuli with regard to the characteristics of the emotion eliciting situation, described in terms of 8 emotion appraisal dimensions (novelty, intrinsic pleasantness, goal conduciveness, urgency, power, self- and other-responsibility, and norm compatibility). First, results showed that the inter-rater reliability was acceptable for all scales except responsibility. Second, the perceived appraisal profiles for the different vocal expressions were generally in accord with predictions based on appraisal theory. Finally, listeners' appraisal ratings on each scale were significantly correlated with several acoustic characteristics. The results show that listeners can reliably infer several aspects of emotion-eliciting situations from vocal affect expressions, and thus suggest that vocal affect expressions may carry cognitive representational information.

1. Background and aims

Recognition of affect in speech is becoming one of the key disciplines within the field of affective computing, and it is also of great relevance to both basic and applied psychological research on emotion (e.g., Elfenbein & Ambady, 2002; Laukka, 2008; Scherer, 2003: Zeng et al., 2009). Much progress has been made in the last decade but fundamental questions like "What affects/emotions can reliably be communicated through the voice?" and "What affects/emotions are universally recognized from the voice?" remain largely unanswered (e.g., Sauter et al., 2010; Simon-Thomas et al., 2009). The first aim of this paper is to introduce the *Vocal Expressions of Nineteen Emotions across Cultures* (VENEC) corpus. This corpus was developed for the purpose of conducting psychological research on nonverbal affect communication with the goal of going some way toward answering the questions above.

Another unsolved question in research on recognition of affect in speech is what methodology is best suited for the annotation of the emotional speech material (Cowie, 2009). The two most common approaches utilized in prior research have been the use of categorical descriptions (e.g., anger, fear, joy, and sadness) combined with forced-choice methodology, or the use of dimensional ratings (most commonly activation and valence). However, both approaches can be criticized for not capturing the full spectrum of nuances that may be present in listeners' perceptions of the emotional content of speech. Our second aim is therefore to present results from initial evaluations of parts of the VENEC corpus using a novel method of annotation, namely letting the listeners rate the stimuli on scales describing emotional appraisal dimensions.

Vocal expressions are thought to convey information about speakers' affective states, but may also reflect the antecedent cognitive appraisal processes that produced the affective states in the speaker (e.g., Johnstone, van Reekum, & Scherer, 2001). It has further been speculated that if vocal expressions do contain information about the emotion-antecedent evaluation processes, this "should allow the listener to reconstruct the major features of the emotion-producing event in its effect on the speaker" (Scherer, 1988, p. 94). Very few previous studies have explored the perception of appraisal dimensions from expressive behavior (Devillers et al, 2006; Scherer & Grandjean, 2008). In this paper, we present the first study that directly explores the amount of information about the emotion-eliciting situation that can be perceived from vocal affect expressions.

2. Methods

2.1 The VENEC corpus

One-hundred professional actors from 5 English speaking cultures (USA, India, Kenya, Singapore, and Australia; 20 from each culture; 50% women; ages = 18-30 years) were recruited to provide the stimuli. They were instructed to vocally express 19 emotions (affection, amusement, anger, contempt, disgust, distress, fear, guilt, happiness, interest, sexual lust, peacefulness, pride, relief, sadness, shame, negative surprise, positive surprise, and emotionally neutral expressions), each with 3 levels of intensity (below average, moderately high, and very high). The particular emotions that are included in various taxonomies of emotion are derived from the particular research methods that were used to obtain the taxonomies (e.g., studies of facial expressions, appraisal dimensions, animal models of physiology, studies on emotion vocabulary), and therefore different taxonomies often include different emotion terms. Our selection of emotion terms is a sample from all various ways of obtaining knowledge about emotions, and presents a comprehensive list that also includes equally many positive and negative emotions.

The actors were provided with scenarios describing typical situations, in which each emotion may be elicited based on current emotion research (e.g., Ellsworth & Scherer, 2003; Lazarus, 1991; Ortony, Clore, & Collins, 1988), and were instructed to try to enact finding themselves in these situations. They were told to try to remember similar situations that they had experienced personally and that had evoked the specified emotions, and if possible try to put themselves into the same emotional state of mind. The actors were further instructed to try to express the emotions as convincingly as possible, but without using overtly stereotypical expressions. The verbal material consisted of short phrases with emotionally neutral content (e.g., "Let me tell you something) and was the same across all expressions. Additionally, each actor also provided a longer paragraph of neutral text, and a subset of the actors provided nonlinguistic vocalizations expressing the above emotions with medium intensity. In all, the corpus includes approximately 6500 stimuli and, for each emotion category, contains a great variety of expressive styles due to speaker, culture, and emotion intensity effects. The whole corpus has further been carefully analyzed with regard to around 70 acoustic cues related to pitch, intensity, voice quality, and durational characteristics using the *Praat* (Boersma & Weenink, 2008) acoustic analysis software (Laukka et al., in press).

2.2 Appraisal rating study

For this study, we used a selection of 300 stimuli from the VENEC corpus, wherein 20 American actors expressed 15 emotions (amusement, anger, contempt, disgust, distress, fear, guilt, happiness, negative surprise, pride, positive surprise, relief, sadness, serenity, and shame) with moderate emotion intensity. Twelve American judges then rated each stimulus with regard to the characteristics of the emotion eliciting situation, described in terms of 8 emotion appraisal dimensions (see below).

The ratings were done separately for each emotion appraisal scale, and the judges were instructed to answer the following questions about the emotion-eliciting event: "NOVELTY - How suddenly and abruptly did the event occur? (not at all sudden/abrupt = 1; very sudden/abrupt = 5): INTRINSIC PLEASANTNESS - How pleasant is the event? (very unpleasant = 1; very pleasant = 5): GOAL CONDUCIVENESS - Did the event help the person to reach a goal or satisfy a need? (prevented a goal/need = 1; helped a goal/need = 5): URGENCY - Did the person need to respond to the event urgently (not at all urgent = 1; very urgent = 5): POWER - Could the outcome of the situation be modified by appropriate human action? (expressor powerless = 1; expressor powerful = 5): SELF-RESPONSIBILITY - Was the speaker responsible for the event? (no personal responsibility = 1; great deal of personal responsibility = 5): OTHER-RESPONSIBILITY - Was another person responsible for the event? (no responsibility by others = 1; great deal of responsibility by others = 5): NORM COMPATIBILITY - Do you think that the event was compatible with the speakers norms? (violated own norms = 1; very consistent with own norms = 5)".

3. Results

First, we wanted to see if the listeners were able to rate the stimuli on appraisal scales in a consistent fashion. The inter-rater reliabilities turned out to be good for all scales, with the exception of self- and other-responsibility, with total reliabilities (Spearman-Brown) ranging between 0.81 – 0.87.

Second, the listeners' appraisal ratings varied systematically as a function of intended emotion (as evidenced by significant effects of intended emotion in repeated measures ANOVAs conducted separately for the listeners' mean ratings on each appraisal dimension). The listeners' mean appraisal ratings as a function of intended emotion are shown in Figure 1. The perceived appraisal profiles for the different intended emotions were largely in accord with predictions based on appraisal theory. For example, anger expressions received high ratings on novelty, urgency, and power, and low ratings on pleasantness, goal conduciveness, and norm compatibility, whereas expressions of amusement received high ratings on pleasantness and norm compatibility, and low ratings on urgency. Also, shame expressions received low ratings on power, norm compatibility, goal conduciveness, and pleasantness. In all, over 90% of the effects of intended emotion were in the predicted direction.

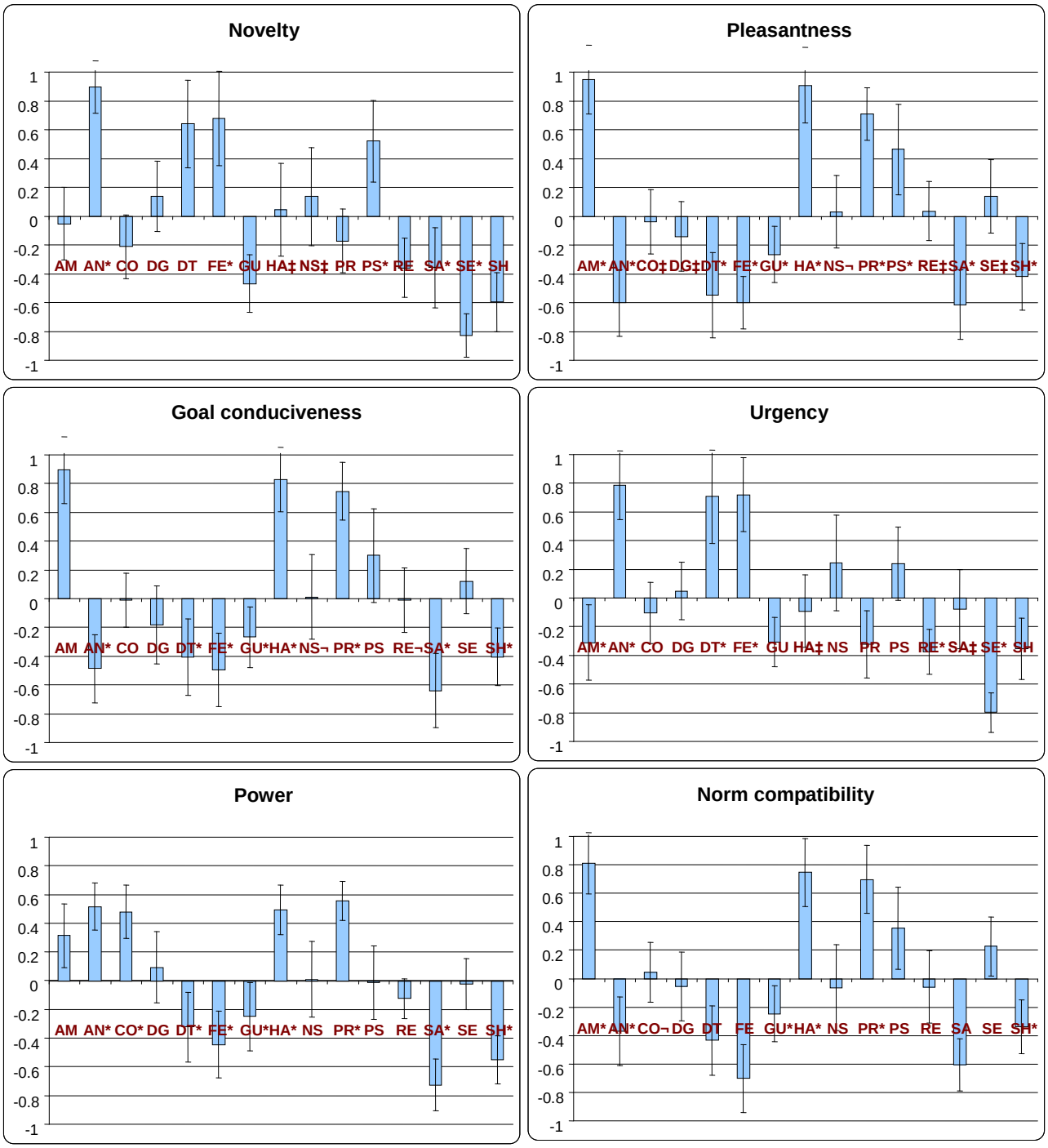

Figure 1: Listeners' mean appraisal ratings (*z*-scores) as a function of intended emotion. Bars indicate 95% confidence intervals. * Result in accordance with theoretical predictions; ‡ Result in predicted direction, but not statistically significant; ¬ Result goes against predictions. Note that we did not have predictions for all emotions/appraisals. Am = Amusement, An = Anger, Co = Contempt, Dg = Disgust, Dt = Distress, Fe = Fear, Gu = Guilt, Ha = Happiness, Ns = Negative surprise, Pr = Pride, Ps = Positive surprise, Re = Relief, Sa = Sadness, Se = Serenity, Sh = Shame.

Acoustic cue	Novelty	Pleasantness	Goal conduciveness	Urgency	Power	Norm compatibility
F0 mean	0.44 ***	0.01 ns	0.01 ns	0.39 ***	-0.18 **	-0.02 ns
F0 sd	0.31 ***	0.16 **	0.17 **	0.16 **	0.21 ***	0.13 *
F0 Q1	0.27 ***	-0.04 ns	-0.06 ns	0.27 ***	-0.28 ***	-0.07 ns
F0 Q5	0.43 ***	0.11 ns	0.09 ns	0.32 ***	-0.05 ns	-0.07 ns
F0 fracrise	0.29 ***	0.11 ns	0.12 *	0.23 ***	-0.04 ns	0.07 ns
Intensity M	0.69 ***	0.11 ns	0.11 ns	0.58 ***	0.26 ***	0.08 ns
H1-A3	-0.26 ***	-0.17 **	-0.17 **	-0.14 *	-0.21 ***	-0.14 *
HF-500	0.34 ***	0.14 *	0.13 *	0.23 ***	0.31 ***	0.15 *
F1 bw	-0.10 ns	-0.18 **	-0.21 ***	-0.03 ns	-0.20 ***	-0.23 ***
% silence	-0.26 ***	-0.37 ***	-0.33 ***	-0.16 **	-0.33 ***	-0.34 ***
Duration	-0.37 ***	-0.10 ns	-0.9 ns	-0.37 ***	-0.14 *	-0.04 ns

Table 1: Correlations between selected acoustic cues and listeners' mean appraisal ratings. F0 mean = mean pitch; F0 sd = pitch variability; F0 Q1 = F0 first quantile; F0 Q6 = F0 fifth quantile; F0 fracrise = proportion of frames with F0 rise; Intensity M = mean intensity; H1-A3 = F0 amplitude - formant 3 amplitude; HF-500 = proportion of high vs. low frequency spectral energy (cutoff: 500Hz); F1bw = median of first formant bandwidth; % silence = proportion of silence in the speech signal; Duration = total duration of the utterance. * $p < .05$; ** $p < .01$; *** $p < .001$. $N = 300$.

Third, listeners' appraisal ratings on each scale were significantly correlated with several acoustic characteristics (see Table 1 for a selection of acoustic cues). Most correlations were small to medium, but give indications of which acoustic cues the listeners utilized in order to make their inferences about the emotion-eliciting situations. For example, high ratings of novelty and urgency were associated with high pitch, high voice intensity, much high-frequency energy, and fast speech rate. High ratings of pleasantness and goal conduciveness, in turn, were associated with large pitch variability and a narrowed bandwidth of the first formant.

4. Discussion

In this paper we presented a new corpus of cross-cultural vocal affect expressions – VENEC – together with a novel way of annotating the perceived emotional content of vocal expressions. Results first showed that the inter-rater reliability was acceptable for all emotion appraisal scales except responsibility. Second, the perceived appraisal profiles for the different vocal expressions were generally in accord with predictions based on appraisal theory. Finally, listeners' appraisal ratings on each scale were significantly correlated with several acoustic cues reflecting pitch, intensity, voice quality, and temporal characteristics).

The results showed that the listeners could reliably infer several aspects of emotion-eliciting situations from vocal affect expressions, which suggests that vocal affect expressions may carry cognitive representational information (see Scherer, 1988). Furthermore, the direct assessment of the amount of information about emotion-eliciting situations that can be perceived from emotion expressions presents a novel way of studying social cognition that is not tied to particular emotion categories. However, more research is needed to determine which appraisal dimensions can be conveyed through the voice, and what combinations of appraisal dimensions are best suited for annotating emotional speech corpora. Future research should also examine similarities and differences between the acoustical correlates of emotion appraisal dimensions and other dimensional representations of emotion (e.g., activation, valence, potency, and emotion intensity; see Laukka, Juslin, & Bresin, 2005).

Extensive annotation of the VENEC corpus is currently underway, using both within- and cross-cultural listening tests and various annotation methods (e.g., forced-choice categorization, free responses) in addition to the appraisal-rating method described in this paper. The corpus is also used in automatic classification experiments, where we investigate the impact of within- and cross-cultural variation in expressive styles on the acoustic characteristics of various emotions.

5. Acknowledgements

The research was funded by the Swedish Research Council through grants to PL (contract 2006-1360) and the National Science Foundation through grants to HAE. We would like to thank all the numerous people all over the world who have helped us in the collection and annotation of the VENEC corpus so far!

6. References

Boersma, P., & Weenink, D. (2008). *Praat: Doing phonetics by computer* [Computer program]. Retrieved from http://www.praat.org/

Cowie, R. (2009). Perceiving emotion: Towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 3515-3525.

Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, S., & McRorie, M. (2006). Real life emotions in French and English TV video clips: An integrated annotation protocol combining continuous and discrete approaches. In *Proc. LREC 2006,* pp. 1105-1110.

Elfenbein, H.A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128,* 203-235.

Ellsworth, P.C., & Scherer, K.R. (2003). Appraisal processes in emotion. In R.J. Davidson, K.R. Scherer & H.H. Goldsmith (Eds.), *Handbook of affective sciences.* New York: Oxford University Press, pp. 572-595.

Johnstone, T., van Reekum, C.M., & Scherer, K.R. (2001). Vocal correlates of appraisal processes. In K.R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research.* New York: Oxford University Press, pp. 271-284.

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,* San Diego, CA: Plural Publishing, pp. 153-169.

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

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K. (in press). Expression of affect in spontaneous speech: Acoustic correlates, perception, and automatic detection of irritation and resignation. *Computer Speech and Language.*

Lazarus, R.S. (1991). *Emotion and adaptation.* New York: Oxford University Press.

Ortony, A., Clore, G.L., & Collins, A. (1988). *The cognitive structure of emotions.* Cambridge, UK: Cambridge University Press.

Sauter, D.A., Eisner, F., Ekman, P., & Scott, S.K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences, USA, 107,* 2408-2412.

Scherer, K.R. (1988). On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology,* 7, 79-100.

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

Scherer, K.R., & Grandjean, D. (2008). Facial expressions allow inference of both emotions and their components. *Cognition and Emotion, 22,* 789-801.

Simon-Thomas, E.R., Keltner, D.J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion, 9,* 838-846.

Zeng, Z., Pantic, M., Roisman, G.I., & Huang, T.S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31,* 39-58.

Towards Measuring Similarity Between Emotional Corpora

Mátyás Brendel, Riccardo Zaccarelli, Björn Schuller, Laurence Devillers

LIMSI-CNRS
Spoken Language Processing Group
91403 Orsay, France

mbrendel, riccardo, schuller, devil@limsi.fr

Abstract

In this paper we suggest feature selection and Principal Component Analysis as a way to analyze and compare corpora of emotional speech. To this end, a fast improvement of the Sequential Forward Floating Search algorithm is introduced, and subsequently extensive tests are run on a selection of French emotional language resources well suited for a first impression on general applicability. Tools for comparing feature-sets are developed to be able to evaluate the results of feature selection in order to obtain conclusions on the corpora or sub-corpora divided by gender.

1. Introduction

At present, there are various corpora in use by the automatic emotion recognition community, with considerable difference in size, topic and application context. None of them are ideal: they all have their advantages and drawbacks. Consequently, the comparison and possible unification of corpora is an important aid for current research: (Tahon and Laurence Devillers, 2010) for example studies differences in anger across corpora in a very detailed fashion by examining acoustic properties. In this paper we take the same two corpora in order to allow for meaningful comparisons and findings, yet with different methods. We will be using PCA and feature selection to visualize and compare corpora by their compound of most relevant features. This seems reasonable, as it is e. g. known that 'more' acted corpora tend to prefer pitch-based descriptors in comparison to more natural emotional speech, where spectral information typically is expected as 'reliable candidate' of best features.

Feature selection is usually considered as a tool to make machine learning models more efficient. Selecting the best features may improve the quality of the model by avoiding over-training and/or it may increase the speed of computation and reduce memory demands. Feature selection, however, is rarely considered as a tool to characterize or analyze a corpus or measure similarities or differences among corpora.

Sequential Floating Forward Selection (SFFS) was introduced in (Pudil et al., 1994) and is certainly among the most widely used techniques in the field. In this paper, we consider SFFS as the baseline experimental technique and try to improve it by the help of 'set-similarity'. This approach is different to other improvements found in the literature, and it introduces a different dimension to amend feature selection algorithms.

As named above, our intent is to analyze the similarities and differences of corpora or sub-corpora by their compound-structure of highly relevant features. The main motivation of introducing a modified variant of SFFS is the comparably high computation cost of SFFS. This is a clear drawback in the respect of our aim: if a feature-selection-based comparison of several corpora is to be carried out, there will be a clear demand for sufficient speed of processing. By using our proposed algorithm a more extensive analysis is possible as it would be achievable in the same amount of time using 'classical' SFFS.

CINEMO	# POS	# SAD	# ANG	# NEU
# segments	313	364	344	510

Table 1: CINEMO sub-corpus, number of segments for 50 speakers

2. Corpora

The corpora CINEMO and JEMO were already introduced in (Brendel et al., 2010). Here we only give a short description.

2.1. CINEMO

The corpus CINEMO (Rollet et al., 2009) used in this paper consists of 1 532 instances after segmentation of emotional French speech amounting to a total net playtime of 2:13:59 hours. 50 speakers (of 15 to 60 years old) dubbed 27 scenes of 12 movies. A subset of the more consensual segments was chosen for training models for detection of 4 classes (POsitive, SADness, ANGer and NEUtral). The rich annotation of CINEMO was used to build these 4 macro-classes. Table 1 shows the distribution of instances among classes within the considered CINEMO sub-corpus.

2.2. JEMO

The corpus JEMO features 1 135 instances after segmentation of speech recorded from 39 speakers (18 to 60 years old). JEMO is a corpus collected within an emotion detection game. This game used a segmentation tool based on silenced pauses and used a first system of 5 emotions detection (ANGer, FEAr, SADness, POsitive and NEUtral) and a system of activation detection (low/high) built on CINEMO data. The corpus recorded was the reaction of the users to the system response.

JEMO	# POS	# SAD	# ANG	# NEU
# segments	316	223	179	416

Table 2: JEMO sub-corpus, number of segments for 39 speakers.

C. & J.	# POS	# SAD	# ANG	# NEU
Male	252	262	267	432
Female	377	325	256	494

Table 3: Female and Male sub-corpora of the unified corpus, # of segments for 38 female and 50 male speakers.

In JEMO speakers generated spontaneous sentences with higher level of expressivity than in CINEMO. The corpus has been annotated by two coders with major and minor emotions. These data were more prototypical than in the corpus CINEMO as very few mixtures of emotions were annotated.
Table 2 shows instance distribution in the JEMO sub-corpus. Table 3 shows instance distribution in the sub-corpora obtained by dividing the unified corpus by gender.

3. Features

In the following we will describe two different feature sets based on two different extraction engines.

3.1. LIMSI features

Each speech segment is passed through spectral (16 MFCCs) and prosodic analysis (pitch, zero-crossing and energy) by the LIMSI extractor. The feature extractor next calculates basic statistical features on voiced parts: min, max, mean, standard deviation, range, median quartile, third quartile, min and max intra range and the mean and standard deviation of the coefficients of least square fitting regression (of each voiced segment); min and max inter range (between voiced segments). Overall, 458 features are thus obtained including further post-processing: 23 for pitch, 51 for energy (from these 22 root mean square energy), 18 zero-crossings and 366 for MFCC1–16.
Table 4 shows the low level descriptors and functionals used in generating the LIMSI features for these experiments.

LLD	**Functionals**
Energy	*moments(2):*
RMS Energy	absolute mean, max
F0	*extremes(3):*
Zero-Crossing-Rate	2 x values, range
MFCC 1–16	*linear regression(2):*
	MSE, slope
	quartiles(2)
	quartile, tquartile

Table 4: *LIMSI features: low-level descriptors and functionals. Abbreviations: root mean square (RMS), Mel Frequency Cepstral Coefficients (MFCC), Mean Absolute/Square Error (MAE/MSE). Note that not all combinations are used.*

LLD	**Functionals**
(δ) RMS Energy	*moments (4):*
(δ) Log-Frame-Energy	absolute mean, std. deviation
(δ) Voicing Probability	kurtosis, skewness
(δ) F0	*extremes (5):*
(δ) F0 envelope	2 $\times$ values, 2 $\times$ position, range
(δ) Zero-Crossing-Rate	*linear regression (4):*
(δ) MFCC 1–12	offset, slope, MAE, MSE
(δ) LSP Frequency 0–7	*quartiles (6):*
	3 $\times$ quartiles, 3 $\times$ ranges

Table 5: *Acoustic features in openEAR: low-level descriptors and functionals. Abbreviations: Line Spectral Pairs (LSP), Mel Frequency Cepstral Coefficients (MFCC), Mean Absolute/Square Error (MAE/MSE).*

3.2. openEAR features

To introduce sufficient variance in our experimentation and not base our findings solemnly on one feature extractor, we use the same openEAR toolkit's (Eyben et al., 2009) *"base"* set as used in (Schuller et al., 2010): 988 features – a slight extension over the set provided for the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009) – based on 19 functionals of 26 acoustic low-level descriptors (LLD, smoothed by simple moving average) and corresponding first order delta regression coefficients as depicted in Table 5.

4. Using PCA for visualizing corpora

To illustrate the distribution of the corpus, the mean of each feature was computed per speaker and per class. In order to be able to display the speaker-means the most important principal components were computed with Weka (Witten and Frank, 2005) and the first two components are shown.
Figure 1 shows the speaker-means of CINEMO with 458 LIMSI features in the first two dimensions of its PCA space.
Figure 2 shows the speaker-means of JEMO with 458 LIMSI features in the first two dimensions of its PCA space.
Figure 3 shows the speaker-means of CINEMO with openEAR features in the first two dimensions of its PCA space.
Figure 4 shows the speaker-means of JEMO with openEAR features in the first two dimensions of its PCA space.
Comparing these figures one can see that the LIMSI features form an elongated shape, especially in JEMO, while with openEAR the shapes are rounder, which seems to be better. Although classes are intertwingled in all the cases, the openEAR figures show more separation of the classes, especially on JEMO. Consequently, we can expect better results with JEMO and with openEAR features. This will be confirmed in the following sections.
PCA is a dimension-reduction technique suited to display instances of a high-dimensional space in a lower dimensional one. However, the principal components are not easily interpretable for humans. It seems worth, though, to select 2 important features and compare the two sub-corpora in this two dimensional and interpretable space.
Figures 5 and 6 show the classes ANG and POS of CINEMO and JEMO in the feature-space of MeanEnergy and MeanPitch. In both cases we can see some differences between

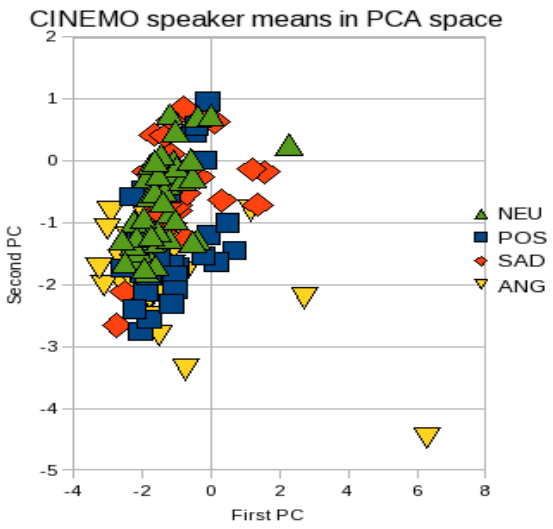

Figure 1: Speaker-means of CINEMO with 458 LIMSI features in the two most important dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

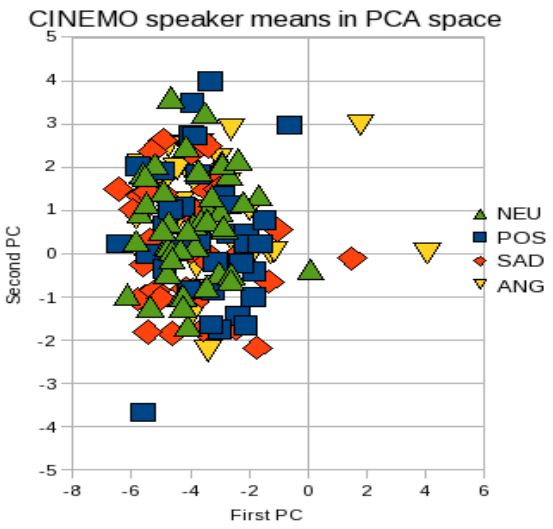

Figure 3: Speaker-means of CINEMO with openEAR features in the two most important dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

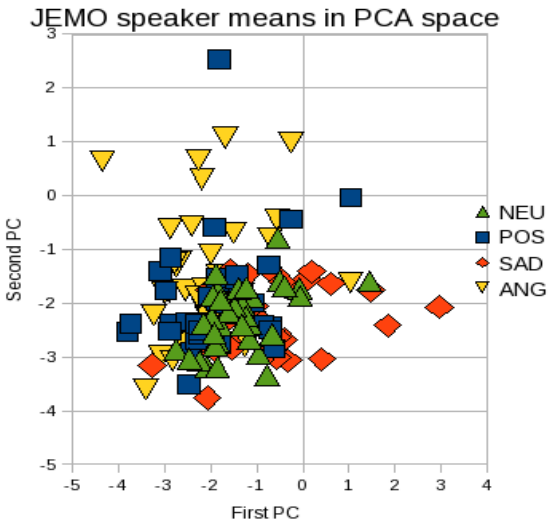

Figure 2: Speaker-means of JEMO with 458 LIMSI features in the first two dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

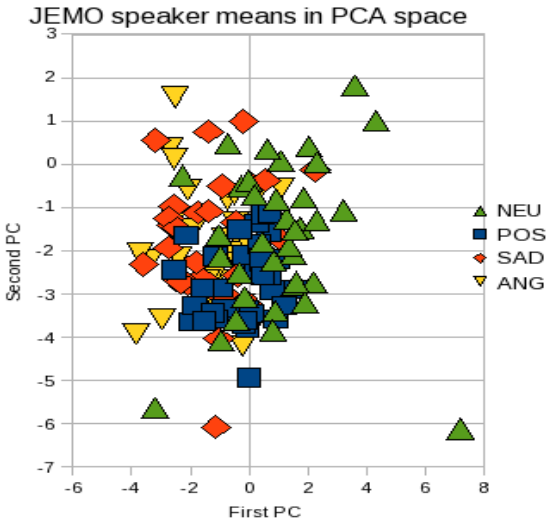

Figure 4: Speaker-means of JEMO with openEAR features in the first two dimensions of its 2D PCA space. Note that some data of all the classes is masked by the classes NEU and POS in the center.

the sub-corpora. Anger in JEMO often contains more energy and higher pitch than in CINEMO. Also in the case of POS of JEMO sometimes higher energetic levels are observed. Note that the coordinates of the two figures are not the same, i. e. the energy related to POSitive is usually lower than that related to ANGer.

5. Using feature selection for measuring differences between corpora

A good introduction to feature selection can be found in (Guyon and Elisseeff, 2003). Methods are generally divided into two larger groups: filter-based and wrapper methods. Filter-based variable ranking is usually computationally affordable, since often only a simple scoring function is computed. However, it usually can not take into account the interaction or correlation between features. At the same time even 'weak' features may add considerably in a compound and should thus not be discarded by choosing only individually high ranked candidates. Wrappers utilize a data-driven learnt classifier's minimal error as target function – consequently they are time-consuming once more complex algorithms are chosen. Usually, one would like to have the later target classifier also employed in the selection process to avoid biases. In this paper we provide results with a wrapper method, namely Sequential Forward Float-

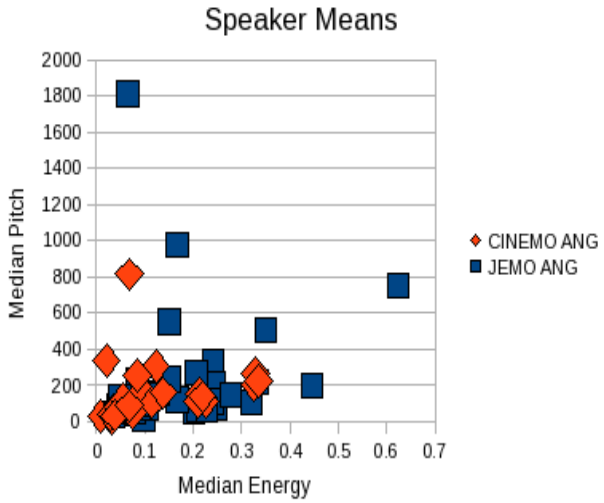

Figure 5: Speaker-means of class ANG in JEMO and CINEMO in the feature-space of MeanEnergy and MeanPitch.

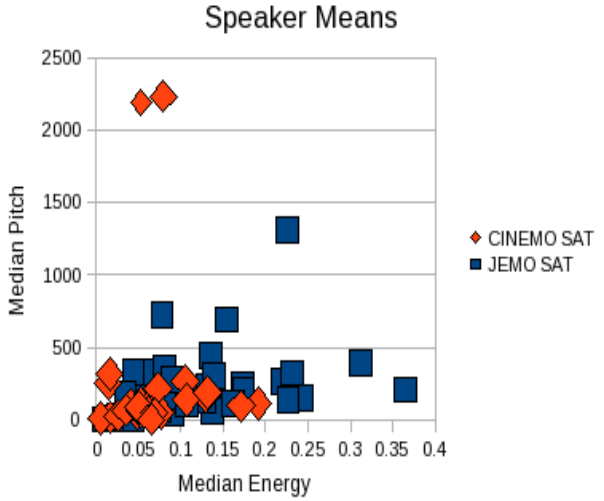

Figure 6: Speaker-means of class POS in JEMO and CINEMO in the feature-space of MeanEnergy and MeanPitch.

ing Search (SFFS) and alternatives that are computationally less 'expensive'.

5.1. Improving feature selection with set-similarity based heuristics

Let $Y = \{y_i : 1 \leq i \leq D\}$ denote a set of available features and $X = \{x_i : 1 \leq i \leq k, x_i \in Y\}$ a subset of features. The named wrapper-methods do feature selection by running a real test on the subset X. In our case this test is always a 10-fold speaker independent (SpI) cross-validation (CV). As described, the worth is evaluated by a measure – in our case recognition rate – and the result of this evaluation is denoted by $J(X)$, i. e. J is our criterion function.

Feature selection can usually be considered as a tree-search, especially when the branch and bound method (B&B) is considered (Somol et al., 2004). Since we consider forward selection, our tree is reverted compared to B&B: the root node represents the empty set and the child nodes of a parent represent all possible extensions of the parent's feature-set with one new feature. The root node has D children, and the number of children is exactly $D - k$ on level k.

As in (Somol et al., 1999), the forward step – the Sequential Forward Search (SFS) – is similar to a breadth-first search on this tree with the rule that on each level only the optimal child is selected while all others are pruned. In addition, there is a backward step, in which a feature is removed, if a better feature-set is obtained measured by the defined optimization criterion than the so far optimal one on that level.

Feature selection can also be considered as a global optimization task, and the forward step can be considered as exploitation and the backward step as exploration. In global optimization the balance of exploration and exploitation is important: without the backward step, SFFS would be too greedy: it would too easily be stuck in local optimums. However, what we will show in the ongoing is that the backward step is strong enough to make the forward step more greedy. At a certain level k, SFFS tests all the $D-k$ nodes and takes the best. A more greedy algorithm would be more similar to a depth-first search, meaning that we take the first child with a positive gain. This way exploitation is made faster, which however, increases the danger to stick to a local optimum. Nevertheless, our claim is that in our field of application, the backward step is strong enough to handle this.

A further improvement was made to order the new feature candidates according to the expectation of their significance. To estimate the gain achievable by adding a certain feature x to the set of features X, we use the known history of our search tree: we take the most similar case, when x was added to a feature set X'. We will call the gain for x the significance of x: $S(x, X)$, which is similar to the notation of (Somol et al., 1999). The estimated significance is denoted as $S'(x, X)$.

Similarity of two sets may be measured in many ways, one of the most frequently used measures is Jaccard similarity, which we decided for. This means that estimated significance is computed as follows:

$$S(x, X) = J(X^* \cup \{x\}) - J(X^*) : \\ X^* = argmax(Jaccard(X, X')) \quad (1)$$

where $Jaccard()$ is the Jaccard set similarity measure. Instead of $argmax$, other functions may be used, like for example a weighted sum with exponentially decaying weights. In the backward step of SFFS we keep the breadth-first manner of the algorithm, since a strong exploration is needed to avoid local minima. However, not a full breadth search is done, only a certain percentage p of the ordered candidates are tested. Candidates are ordered in increasing order of estimated significance, so that we try to remove first the most insignificant features. We applied p=20 % of breadth search, which proved to be sufficient in our case.

We name our introduced method "SFFS with Set-Similarity Heuristics" (SFFS-SSH). This proposed algorithm has only one parameter to be tuned: p. Note that the computational overhead of our heuristics is negligible compared to a 10-fold CV test on a corpus. Thus, the running time of the

# it. / RR	SFFS	SFFS-SSH	all feat.
LIMSI	28 382 / 53.4 %	6 394 / 52.4 %	– / 54.6 %
openEAR	28 126 / 58.5 %	4 742 / 58.8 %	– / 59.6 %

Table 6: Number of iterations and Recognition Rate (RR) using the best 24 selected features on the united CINEMO and JEMO corpus with conventional SFFS in comparison to our suggested efficient modification (SFFS-SSH).

RR	24 L.	24 O.	all L.	all O.
CINEMO	49.3 %	56.6 %	48.5 %	53.8 %
JEMO	63.7 %	67.9 %	61.6 %	64.2 %
Female	64.6 %	64.5 %	59.5 %	62.6 %
Male	53.0 %	56.6 %	49.5 %	57.0 %

Table 7: Recognition Rate (RR) with the best 24 selected features, and all features. Abbreviations: openEAR (O.), LIMSI-features (L.)

algorithm depends mainly on the number of iterations, which we expect to decrease significantly.

In the following experiments we trained the data set using LIBSVM (Chih-Chung Chang and Lin, 2001) with a radial basis function kernel. As stated earlier, we use 10-fold speaker independent cross-validation, designed in our lab. In short this means that speakers are divided into folds, instead of partitioning merely taking instances without speaker assignment into account, thus maintaining speaker independence, while being able to run a 10-fold CV with all its benefits as being able to use a complete (sparse) data set for testing and introduce variance in the evaluative runs.

Table 6 shows the results with a fixed number of 24 features. As can be seen, selecting 24 features does not improve the result. The explanation for this is that the united corpus is sufficiently large to avoid over-training. It can further be seen that SFFS-SSH provides similar good feature-selection at considerably lower number of iterations. Thus the first test of our method was successful, which is why we will exclusively apply SFFS-SSH in the ongoing.

Table 7 shows the recognition rate (RR) for the various sub-corpora with SFFS-SSH with the number of features fixed to 24.

There is a more significant difference in RR between CINEMO and JEMO than between the female and male sub-corpora for both libraries (openEAR and LIMSI) and also for the selected features and the total set of features.

Table 8 next shows the recognition rate (RR) for the various sub-corpora with SFFS-SSH with the optimal number of features.

RR / # features	LIMSI	Openear
CINEMO	55.7 % / 29	58.2 % / 36
JEMO	65.8 % / 43	72.2 % / 43

Table 8: Recognition Rate (RR) and number of features of the best selected feature set

LIMSI	MFCC	Pitch	Energy	ZCR
CINEMO	21	0	2	1
JEMO	21	0	2	1
Female	17	3	3	1
Male	17	5	2	0

Table 9: Frequency of different feature groups in the 24 selected LIMSI features.

openEAR	MFCC	Pitch	Energy	Zrc
CINEMO	12	0	6	1
JEMO	11	5	2	0
Female	10	3	3	4
Male	6	1	8	0

Table 10: Frequency of different feature groups in the 24 selected openEAR features. Note that if features are missing to sum up to 24, they are of other kind than the considered ones.

As seen in the table, some percent of improvement can be achieved with a significantly higher number of features. Interestingly, the optimal number of features is higher for JEMO with both feature-sets.

5.2. Ratio of feature groups in selected feature sets

Having established an efficient method for feature selection we will next consider how it can be used for our primary aim in this paper: in order to measure the differences between sub-corpora, as a first attempt we have computed the ratio of different feature groups for LIMSI and openEAR after feature selection.

Tables 9 and 10 show the number of features in the different features groups. Since the number of features is constantly 24, these correspond to ratios. There are considerably less MFCC features used in openEAR in overall ratio, but more corresponding to energy and other low-level descriptors, not grouped here (cf. table 5).

The differences between female-male seems to be larger than between CINEMO and JEMO. This contradicts the difference in RR. Consequently, these numbers are important but not detailed enough. We thus will next investigate further measures.

Since CINEMO and JEMO in table 9 have exactly the same ratios for each feature group, we repeated this experiment with 48 features. This naturally demands for longer computation times, as SFFS is a forward selection. Thus, it would have been desirable to reveal differences already at a low dimension of the selected feature space. For quantitative illustration we consider this experiment with the LIMSI set on the CINEMO and JEMO corpora. Results are shown in table 11.

Visibly, there is a slight difference in this case compared to the smaller target set size, but results across corpora are very similar. Comparing table 9 to table 11, we can only see that more energy features have been selected, which likely indicates their lower relevance, though used, if more features are to be selected. Recognition rate is 64.6 % for

LIMSI	MFCC	Pitch	Energy	ZCR
CINEMO	32	1	13	2
JEMO	30	1	16	1

Table 11: Number of different feature groups in the 84 selected LIMSI features.

Similarity of features	LIMSI	openEAR
CINEMO–JEMO	0.5983	0.5903
Female–Male	0.4907	0.4255

Table 12: Correlation-based similarity of the selected feature-sets.

JEMO and 51.1 % for CINEMO, which resembles a slight improvement.
This extended experiment did not bring us further – it just confirmed previous results: we consider more features and by that obtain slightly improved results in terms of recognition rate, however, we still need tools for more detailed analysis of the features.

5.3. Correlation based similarity of feature-sets

Having two feature sets from the same total set of features, one would like to compare the two sets. To list the features selected appears to have less practical applicability, since there might be different features, which are similar. For the same reason, a simple Jackard-similarity is also not sufficient. Instead, measures of describing and comparing feature-sets have to be developed.
One way to measure similarity of corresponding features is the cross-correlation matrix. There is no way to define a cross-correlation between the same features of the different sub-corpora, like CINEMO and JEMO or female and male, since the instances are independent. What can be done, though, is to compute a cross-correlation for different features over the united corpus, i. e. for CINEMO and JEMO together. Moreover, we can define the similarity of the selected feature-sets based on this. The similarity of feature sets F and F' is computed as follows:

$$sim(F, F') = aver\{f \in F : min\{corr(f, f') : f' \in F'\} \quad (2)$$

Where "aver" is the average and "corr" the correlation computed on the entire corpus. Since the measure "sim" is asymmetrical, the average of $sim(F, F')$ and $sim(F', F)$ was taken. Note that this measure would not have any sense over the totals of features (it would be 1) – it is only reasonable in combination with feature selection.
In table 12 the similarity of sub-corpora can be seen measured by the similarity of the 24 selected features. As can be seen, the similarity of CINEMO and JEMO in terms of feature sets is higher than of female to male. This confirms our finding in recognition rate.

5.4. Rank based similarity of feature-sets

It is not straight forward to derive an individual feature's relevance in the resulting feature-set. To obtain a sharpened picture on this, one can use the first iteration of SFFS as a feature ranking, as it considers each individual. Note that the first iteration of SFFS and our SFFS-SSH is identical. The rank of the feature is the result obtained using only that one feature. One can do this not only for the selected features but for the total set of features, i. e. 458 for LIMSI features and 988 for openEAR in our case.

Difference in feature ranks	LIMSI	openEAR
CINEMO-JEMO	0.17	0.19
Female-Male	0.07	0.09

Table 13: Highets differences in feature ranks.

Table 13 shows the highest difference in feature ranks for the same features. This measure shows that CINEMO and JEMO is more different in this aspect than the female and male sub-corpora. There is a consistently higher difference for openEAR, the reason might be that results for openEAR are better and the number of features is higher, which allows for the maximum to be higher. The measure based on feature ranks by that confirms the measure based on feature-groups.

6. Conclusions

We have seen four measures connected to feature selection for measuring similarity of sub-corpora: similarity measures based on recognition rate, groups of features, correlation and feature-ranks. They support however two different kinds of results: for recognition rate and correlation, the difference between female and male – which was considered as comparative anchor – is higher than between CINEMO and JEMO. On the other hand, measured in feature-groups and feature-ranks, the difference between CINEMO and JEMO is higher than the difference between female and male. Our result indicates that several measures have to be used. The four measures seem also to show that there might be at least two aspects of difference: in one aspect female and male are more different, in another the CINEMO and JEMO sub-corpora are.

7. Future Work

For dimension reduction, future work may test other techniques, especially so called cluster-preserving or similarity-preserving transformations.
In the future several more measures shall be developed to measure similarity of feature sets and corpora. Similarity of feature sets and the importance of features might be represented in a tree-like structure, which corresponds to the tree-structure of feature selection and propagates some measures of importance and similarity. This structure and measure is however complex.
Our SFFS-SSH algorithm can also be further developed. For example, instead of Jaccard-similarity a correlation based similarity measure may be used not only after feature selection but already within feature selection.
The correlation-based similarity measure can also be improved: for each feature currently we take only into account the most similar feature. A more complex measure would add together all the similar features with a decreasing weight.

8. References

M. Brendel, R. Zaccarelli, and L. Devillers. 2010. Building a system for emotions detection from speech to control an affective avatar. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

C.-C. Chih-Chung Chang and C.-J. Lin, 2001. *LIBSVM: a library for support vector machines*.

F. Eyben, M. Wöllmer, and B. Schuller. 2009. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. ACII*. IEEE.

I. Guyon and A. Elisseeff. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

P. Pudil, J. Novovičová, and J. Kittler. 1994. Floating search methods in feature selection. *Pattern Recogn. Lett.*, 15(11):1119–1125.

N. Rollet, A. Delaborde, and L. Devillers. 2009. Protocol cinemo: The use of fiction for collecting emotional data in naturalistic controlled oriented context,. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction, 2009*.

B. Schuller, S. Steidl, and A. Batliner. 2009. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*, pages 312–315, Brighton, UK. ISCA.

B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers. 2010. Cinemo a french spoken language resource for complex emotions: Facts and baselines. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

P. Somol, P. Pudil, J. Novovičová, and P. Paclík. 1999. Adaptive floating search methods in feature selection. *Pattern Recogn. Lett.*, 20(11-13):1157–1163.

P. Somol, P. Pudil, F. J. Ferri, and J. Kittler. 2004. Fast branch & bound algorithms for optimal feature selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(7):900–912.

M. Tahon and L. Laurence Devillers. 2010. Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Proceedings of the Fifth International Conference on Speech Prosody, 2010*.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database

Michel F. Valstar, Maja Pantic

Imperial College London / Twente University
Department of Computing / EEMCS
180 Queen's Gate / Drienerlolaan 5
London / Twente
Michel.Valstar@imperial.ac.uk, M.Pantic@imperial.ac.uk

Abstract

We have acquired a set of audio-visual recordings of induced emotions. A collage of comedy clips and clips of disgusting content were shown to a number of participants, who displayed mostly expressions of disgust, happiness, and surprise in response. While displays of induced emotions may differ from those shown in everyday life in aspects such as the frequency with which they occur, they are regarded as highly naturalistic and spontaneous. We recorded 25 participants for approximately 5 minutes each. This collection of recordings has been added to the MMI Facial Expression Database, an online accessible, easily searchable resource that is freely available to the scientific community.

1. Introduction

A key goal in Automatic Human Behaviour Analysis is "really natural language processing" (Cowie and Schröder, 2005), which endows machines with the ability to speak to human users in the same way that a human would speak to another person. Achieving that goal includes finding ways to make machines understand the non-verbal signals that humans send and which are part and parcel of human conversation. Emotion is one signal that Automatic Human Behaviour Analysis scientists have focused on for approximately thirty years already. Automatic detection of emotions has been investigated both from audio, video, and recently also by fusing the audio and video modalities (see (Zeng et al., 2009) for an overview).

Following Darwin, discrete emotion theorists propose the existence of six or more basic emotions that are universally displayed and recognised (Darwin, 1872). These emotions are Anger, Disgust, Fear, Happiness, Sadness, and Surprise. Data from both modern Western and traditional societies suggest that non-verbal communicative signals involved in these basic emotions are displayed and recognised cross-culturally (Keltner and Ekman, 2000). While the basic emotions do not occur very frequently in normal human-human interactions, when they do occur they convey a very strong message to someone's surroundings.

A major issue hindering new developments in the area of Automatic Human Behaviour Analysis in general, and affect recognition in particular, is the lack of databases with natural displays of behaviour and affect. While a number of publicly available benchmark databases with posed displays of the six basic emotions exist, and are well studied (Pantic et al., 2005; Lyons et al., 1998; Kanade et al., 2000), there is no equivalent of this for spontaneous basic emotions.

There do exist a few databases that contain spontaneous emotive content. Table 1 gives an overview of them. In the second to fourth columns of the table, we list how many people were recorded, the total duration of the dataset, and the bandwidth with which the data was recorded. In the 'Availability' column we list whether the database is freely available to anyone (*Public*), freely available to the academic scientific community (*Scientific*), or not available at all (*Private*). In the 'Online' column we list whether the data is an on-line repository, or whether the data is distributed by traditional mail. Finally, in the last column we indicate whether there is an online search option to the database, which would allow researchers to select and download exactly the set of data they require.

Two databases that have been used recently in studies of automatic human behaviour analysis on spontaneous data are the RU-FACS database (Bartlett et al., 2006) and the DS-118 database (Rosenberg et al., 1998). RU-FACS was recorded at Rutgers University. A hundred people participated in false opinion paradigm interviews conducted by retired law enforcement agents. Participants had to either lie or tell the truth about a social/political issue. If participants chose to lie about their opinions, they would gain 50 US Dollars if they convinced the interviewers of their views, and were told that they would have to fill out a long and boring questionnaire if they failed. This raised the stakes of lying, and thus elicited stronger and more natural expressive behaviour.

The DS-118 dataset has been collected to study facial expression in patients with heart disease. Subjects were 85 men and women with a history of transient myocardial ischemia who were interviewed on two occasions at a 4-month interval. Spontaneous facial expressions were recorded during a clinical interview that elicited spontaneous reactions related to disgust, contempt, and other negative emotions as well as smiles.

Unfortunately, these two databases are not freely available to the scientific community. This makes it impossible to reproduce results of automatic behaviour analysis that are tested solely on these databases. Three databases containing displays of the basic emotions that *are* freely available to the academic community are the SAL (Douglas-

Table 1: Overview of databases with spontaneous emotive content.

Database	Participants	Duration	Video Bandwidth	Audio Bandwidth	Availability	Online	Searchable
DS-118 (Rosenberg et al., 1998)	100	4:10:00	unknown	unknown	Private	No	No
MMI-db Part IV & Part V	25	1:32:00	640x480 pixels @ 29Hz	44.1kHz	Scientific	Yes	Yes
RU-FACS (Bartlett et al., 2006)	100	4:10:00	unknown	unknown	Private	No	No
SAL (Douglas-Cowie et al., 2007)	4	4:11:00	352x288 pixels @ 25 Hz	20kHz	Scientific	Yes	No
SEMAINE (McKeown et al., 2010)	20	6:30:41	580x780 pixels @ 49.979 Hz	48kHz	Scientific	Yes	Yes
Spaghetti db (Douglas-Cowie et al., 2007)	3	1:35	352x288 pixels @ 25Hz	Unknown	Scientific	Yes	No
Vera am Mittag db (Grimm et al., 2008)	20	12:00:00	352x288 pixels @ 25Hz	16kHz	Public	No	No

Cowie et al., 2007), SEMAINE (McKeown et al., 2010), and Spaghetti (Douglas-Cowie et al., 2007) databases. Both the SAL and SEMAINE databases record interactions between a user (the experiment participant) and an operator (someone from the experimenters' team). The operators act out one of four prototypic characters: one happy, one sad, one angry and one neutral. This results in emotionally coloured discourse. As shown in table 1 both databases contain a considerable amount of data. However, except for the emotion 'happiness', emotions are mostly displayed in a very subtle manner. While these databases are very suitable for analysing expressions displayed in natural discourse, they are less suitable for training systems that can identify the basic emotions.

The Spaghetti database on the other hand does contain strong basic emotions; mostly of fear, disgust, surprise, and happiness. The database consists of recordings of an experiment where people were asked to feel inside a box that contained a warm bowl of spaghetti. Because the participants didn't know what's in the box, they reacted strongly when their hands touched the spaghetti. The data was released as part of the HUMAINE database (Douglas-Cowie et al., 2007). Unfortunately, it consists of recordings of only three participants, and the total dataset lasts only one minute and 35 seconds. This makes it very hard to train any automatic Human Behaviour Understanding algorithms on this data.

The MMI-Facial Expression database was conceived in 2002 by Maja Pantic, Michel Valstar and Ioannis Patras as a resource for building and evaluating facial expression recognition algorithms (Pantic et al., 2005). Initially the focus of the database was on collecting a large set of AUs, occurring both on their own and in combination, from either videos or high-quality still images. Later data to distinguish the six basic emotions were added, and in this work the addition of spontaneous data is described.

Recording truly spontaneous instances of basic emotion expressions is extremely difficult, because in everyday life the basic emotions aren't shown frequently. However, when they *are* displayed, they convey a very strong message to someone's surroundings, one which should certainly not be ignored by Automatic Human Behaviour Analysis systems. In order to record a truly spontaneous dataset of sufficient size[1] it would thus be necessary to follow and record the participants for a very long duration. Following the participants for a long time would mean the recording setup would need to be compact and mobile. A side effect of this would be that one loses the ability to control the recording conditions. Instead of waiting for the expressions to occur naturally, we decided to induce them by showing the participants a collage of short video clips. The clips were selected to induce the emotions happiness and disgust.

The remainder of this work is structured as follows: section 2. explains how we recorded the data. The recorded data was manually annotated for the six basic emotions and facial muscle actions (FACS Action Units (Ekman et al., 2002)). The annotation details are presented in section 3.. Section 4. describes the existing MMI Facial Expression Database and the place that the new induced emotional recordings described in this work have in it. We provide a summary and closing remarks in section 5..

2. Data acquisition

The goal of the experiments is to record naturalistic audiovisual expressions of basic emotions. Due to ethical concerns, it is very hard to collect such data for the emotions anger, fear, and sadness. Happiness, surprise and disgust on the other hand are easier to induce. We collected data from these three expressions in two different experiments. In the first experiment we showed the participants short clips of cartoons and comedy shows to induce happiness. Images and videos of surgery on humans and humans effected by various diseases were shown to induce disgust. In the second experiment, only the happiness inducing type of clips were shown. In both experiments, the participants showed expressions of surprise too, often mixed with either happiness or disgust.

The images and videos were shown on a computer screen. The participants were sitting in a chair at approximately 1.5 metres distance to the screen. In both experiments, a JVC GR-D23E Mini-DV video camera with integrated stereo microphones was used to record the reactions of the participants. The camera was placed just above the computer screen, ensuring a near-frontal view of the participants' face as long as they face the screen.

During the recording of the first experiment the experimenters were in the room with the participants. As a result the participants engaged in social interactions with the experimenters about the content of the images shown, and what they thought about this. We regarded this as undesirable components of the recordings, as the experimenters influenced the behaviour of the participants. The recordings were manually cut into 383 segments (called Sessions) that contain distinct displays of affective behaviour. To distinguish this set of data from existing datasets in the database, we will refer to this as Part IV of the MMI Facial Expression database (see section 4.).

In the second experiment, the participants would hear the

[1] Sufficient meaning enough recordings to be able to perform studies that can have statistically significant results.

Figure 1: A selection of expressions of disgust and happiness added to the MMI-Facial Expression Database. The first row is taken from Part V of the database and shows expressions of happiness and disgust. The second row shows expressions of happiness and disgust taken from Part IV of the database. The third row shows four frames of a single sequence in which the participant showed an expression of disgust.

sound of the stimuli over headphones instead of over computer speakers, as was the case in experiment 1. This resulted in less noise in the audio signal. Another refinement was that the participants were left in a room on their own while the stimuli were provided. The idea behind this is that the participants would be less socially inhibited to show their emotions if there were no other people around. Also, without interruptions caused by interactions between participants and experimenters, the data can now be used to analyse the changes in behaviour over time. To allow the latter, we chose not to cut the recordings of the second experiment into smaller clips. We will refer to this data as Part V of the MMI Facial Expression database.
In total, 25 participants aged between 20 and 32 years took part in the experiments, 16 in experiment 1 and 9 in experiment 2. Of these, 12 were female and 13 male. Of the female participants, three were European, one was South American, and eight were Asian. Of the men, seven were European, two were South American and four were of Asian background.

3. Annotation of affect, laughter and facial muscle actions

Part IV of the database has been annotated for the six basic emotions and facial muscle actions. Part V of the database has been annotated for voiced and unvoiced laughters.

3.1. Emotion and Action Unit annotation

All Sessions of Part IV were annotated to indicate which of the six basic emotions occurred in each clip. This annotation is valuable for researchers who wish to build automatic basic emotion detection systems. Not all affective states can be categorised into one of the six basic emotions. Unfortunately it seems that all other affective states are culture-dependent.

Instead of directly classifying facial displays of affective states into a finite number of culture-dependent affective states, we could also try to recognise the underlying facial muscle activities. These muscle activations are objective measures of facial expression. These can then be interpreted in terms of (possibly culture-dependent) affective categories such as emotions, attitudes or moods. The Facial Action Coding System (FACS) (Ekman et al., 2002) is the best known and the most commonly used system developed for human observers to describe facial activity in terms of visually observable facial muscle actions (i.e., Action Units, AUs). Using FACS, human observers uniquely decompose a facial expression into one or more of in total 31 AUs that produced the expression in question.

Note that all possible facial expressions can be uniquely described by this 31-dimensional space. That is a rather low-dimensional space, which means that learning a mapping from AUs to affective states requires significantly less space than a mapping directly from images to affective states would need.

All Sessions belonging to Part IV have been FACS AU-coded by a single FACS coding expert. For each Session, the annotation indicates which AUs were present at some time during the clip.

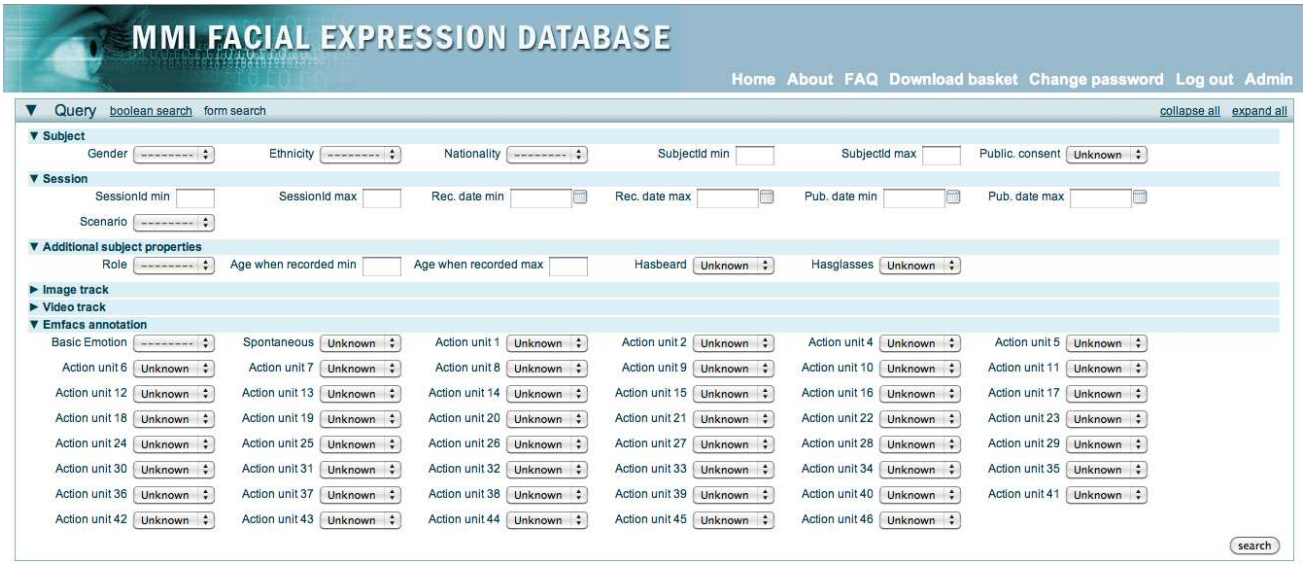

Figure 2: The online search form of the MMI Facial Expression Database.

3.2. Laughter annotation

Laughter is a an event in which a person smiles and produces a typical sound. It is therefore an audio-visual social signal. Laughters have been further divided into two categories: voiced and unvoiced laughters (Bachorowski et al., 2001). To allow studies on the automatic audio-visual analysis of laughter (e.g. (Petridis and Pantic, 2009)), we annotated in all recordings of Part V of the database exactly when voiced and unvoiced laughters events occurred. In our annotation rules, a laughter event was coded as voiced if any part of that laughter event had a voiced component.
Part V consists of nine recordings. In total we, annotated 109 unvoiced and 55 voiced laughters. The average duration of an unvoiced laughter was 1.97 seconds, while a voiced laughter lasted 3.94 seconds on average (exactly twice as long). The laughter annotation was performed using the ELAN annotation tool (Wittenburg et al., 2006).

4. MMI Facial Expression Database

The MMI Facial Expression Database is a continually growing online resource for AU and basic emotion recognition from face video. In the following sections we will describe the database's structure, and how to use its web-interface.

4.1. Database organisation

Within the database, the data is organised in units that we call a *Session*. A Session is part of a *Recording*, which is a single experiment, i.e. all data of a single participant watching all the stimuli. Each Session has one or more single sensor data files associated with it. For the data presented in this paper, this is the audio-visual data stream recorded by the camera. We call these database entries *Tracks*. There are low-quality previews of all tracks on the web-based interface of the database so that users can get an idea of what content each track has, before choosing to download it. The fourth component that makes up the database are the *annotations*. In the case of the MMI Facial Expression database, FACS and basic emotion annotation data is available in so-called EMFACS annotations, and laughter annotations are available in the form of ELAN files.
The first data recorded for the MMI Facial Expression Database were mostly displays of individual AUs. In total, 1767 clips of 20 participants were recorded. Each participant was asked to display all 31 AUs and a number of extra Action Descriptors (ADs, also part of FACS). After all AUs were recorded, the participants were asked to perform two or three affective states (e.g. sleepy, happy, bored). The participants were asked to display every facial action twice, to increase the variability of the dataset. During recording a mirror was placed on a table next to the participant at a 45 degree angle with respect to the camera. This way, a completely synchronised profile view was recorded together with the frontal view. We will refer to this part of the data as Part I of the database. It consists of Sessions 1 to 1767, and is video-only.
The second set of data recorded were posed displays of the six basic emotions. In total, 238 clips of 28 subjects were recorded. Again, all expressions were recorded twice. People who wear glasses were recorded once while wearing their glasses, and once without. This time we focused on obtaining a higher spatial resolution of the face. Therefore we did not use a mirror to obtain a profile view, and we tilted the camera to record the faces in portrait-orientation. We will refer to this part of the data as Part II of the database. It consists of Sessions 1767 to 2004, and is video-only.
Part III, the third set of data recorded consists of high-quality still images. Similar to Part I, the participants were asked to display all AUs and the six basic emotions. In total 484 images of 5 subjects were recorded. Part III consists of Sessions 2401-2884. The acquisition of Parts I-III are described in detail in (Pantic et al., 2005).
0Part IV and Part V are described in detail in section 2. . Part IV consists of Sessions 2005-2388, while Part V consists of Sessions 2895 to 2903.

4.2. Search

The web-based interface of the database has search form functionality that allows the user to collect and preview the exact set of data that he or she needs. Search options include searching on basic emotion, FACS coding, whether the expression was posed or spontaneous, ethnographic variables of the subjects and technical details of the recorded media. Figure 2 shows the search form layout.

One particularly interesting search criterium is the recording scenario type. To describe the different possible scenarios in which emotionally coloured data can be acquired, we divide the possible scenarios in three groups: *acting*, *reacting*, and *interacting*. In this taxonomy, all posed expression databases fall in the category *acting*, while all scenarios in which an emotion is induced by providing the experiment participants with a certain stimulus while measuring their reactions fall in the *reacting* category. The additional datasets described in this work (i.e. Part IV and Part V) are classic examples of *reacting* scenarios. Finally, scenarios in which participants are interacting freely fall in the *interacting* category. Good examples of this are the SAL and SEMAINE databases (Douglas-Cowie et al., 2007; McKeown et al., 2010).

So, to find parts IV and V within the MMI Facial Expression database, one possibility would be to search for the scenario option *reacting*, in the EMFACS section of the search form. Alternatively one could search directly using the Session numbers described above.

4.3. Database availability

The database is freely available to the academic scientific community, and is easily accessible through a web-interface. The url of the database is http://www.mmifacedb.com. Prospective users have to register with that website to create an account. To activate this account, the prospective users need to sign and return an End User License Agreement (EULA). Among other things, the EULA prohibits using the data to train or test commercial products, and it prohibits all military use.

The EULA allows the user to use imagery contained in the database for academic publications and presentations, provided that the participants shown in that particular imagery have specifically allowed this. One can search for this in the search form.

5. Conclusion

We have presented an addition to the MMI-Facial Expression corpus consisting of spontaneous data. Emotions of happiness, disgust, and surprise were induced in 25 participants and their displays/outbursts of affect were recorded on video and audio. This resulted in 1 hour and 32 minutes of data, which is made available in 392 segments called Sessions. Part IV of the data was manually annotated for the six basic emotions, and for FACS Action Units. Part V was annotated for voiced/unvoiced laughters. We believe this dataset can be of great benefit to researchers in the field of Automatic Human Behaviour Understanding.

6. Acknowledgments

This work has been funded in part by the European Community's 7th Framework Programme [FP7/20072013] under the grant agreement no. 231287 (SSPNet). The work of Michel Valstar is further funded in part by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no. 211486 (SEMAINE). The work of Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

7. References

J. A. Bachorowski, M. J. Smoski, and M. J. Owren. 2001. The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110(1):1581–1597.

M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, and J.R. Movellan. 2006. Automatic recognition of facial actions in spontaneous expressions. *Journal of Mutlimedia*, pages 1–14, Oct.

R. Cowie and M. Schröder. 2005. Piecing together the emotion jigsaw. *Machine Learning for Multimodal Interaction*, pages 305–317, Jan.

C. Darwin. 1872. *The Expression of the Emotions in Man and Animals*. John Murray, London.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, Lowry O., M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. 2007. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:488–501, Jan.

P. Ekman, W.V. Friesen, and J.C. Hager. 2002. *Facial Action Coding System*. A Human Face.

M. Grimm, K. Kroschel, and S. Narayanan. 2008. The vera am mittag german audio-visual emotional speech database. *IEEE International Conference on Multimedia and Expo*, pages 865–868.

T. Kanade, J.F. Cohn, and Y. Tian. 2000. Comprehensive database for facial expression analysis. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53.

D. Keltner and P. Ekman. 2000. Facial expression of emotion. In M Lewis and J.M. Haviland-Jones, editors, *Handbook of emotions*, pages 236–249. Guilford Press.

M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. 1998. Coding facial expressions with gabor wavelets. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205.

G. McKeown, M.F. Valstar, M. Pantic, and R. Cowie. 2010. The semaine corpus of emotionally coloured character interactions. *Submitted to ICME*, pages 1–6, Jan.

M. Pantic, M.F. Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo*, pages 317–321.

S. Petridis and M. Pantic. 2009. Is this joke really funny? judging the mirth by audiovisual laughter analysis. In *IEEE International Conference on Multimedia and Expo*, pages 1444 –1447, 28 2009-july 3.

E.L Rosenberg, P. Ekman, and J.A. Blumenthal. 1998. Facial expression and the affective component of cynical hostility in male coronary heart disease patients. *Health Psychology*, 17(4):376–380, Aug.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. *Proceedings of Language Resources and Evaluation Conference*.

Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

Complementing Datasets for Recognition and Analysis of Affect in Speech

Tal Sobol-Shikler
Present Address: Department of Industrial Engineering and Management, Ben-Gurion University of the Negev
Marcus Campus, POB 653, Beer Sheva 84105, Israel
E-mail: stal@bgu.ac.il, tal.shikler@gmail.com

Abstract

The paper presents a framework for an automatic system that infers affective states from their non-verbal expressions in speech. The goal was to infer affective states occurring in real scenarios, i.e. affective states which are common in everyday lives and in human-computer interactions, can occur simultaneously and whose level of expression can change dynamically and asynchronously over time, as well as investigating the generalization to a very large variety of affective states and to other languages. The framework was based on two complementing datasets, Mind Reading and Doors, which were used for the design and validation of the system. The chosen datasets provided data in two languages, by speakers of both genders and of all age groups, actors and non-actors. The data comprised of acted and naturally evoked affective states, with varied text and text repetitions, labeled affective states and multi-modal information, single sentences and sustained interactions, a large variety of affective states and nuances of subtle affective states, in two different languages (English and Hebrew). The paper shortly describes the datasets, their advantages and disadvantages, and how their combination was used in order to achieve the design goals.

1. Introduction

The recognition of affective states from speech has the potential to enhance many human-computer, human-robot and computer-mediated interfaces (Picard, 1997, Zeng *et al*., 2009). Recognition results can be used for analysis of user reactions in order to predict intentions and to generate appropriate response. It can also be used for annotation of speech corpora for synthesis of affective speech. In order to achieve that, the systems designed should be able to infer affective states occurring in real scenarios. The development of such systems entails collection and labeling of speech corpora (Douglas-Cowie *et al.* 2003), development of signal processing and analysis techniques, as well as consolidation of psychological and linguistic analyses of affective states (Cowie *et al.*, 2001, Sobol-Shikler, 2010). In order to apply to a wide variety of applications, the recognition of affect in speech should be able to encompass a wide variety of affective states which occur in real settings, including emotions, attitudes, beliefs, intents, desires, pretending, knowledge and moods. Such systems should also address affective states that occur simultaneously, nuances of expressions and dynamic (and asynchronous) changes of affective states and their expression levels over time. The ability to generalize recognition systems to different speakers and possibly to different languages is also desirable for valid applications. These goals define the requirements from the datasets in use and of the design, implementation and validation of a designed system.

A major challenge in this research area is the lack of conventional, public databases of naturally evoked, labeled affective states, both for single mode and for multi-modal analysis. This shortage requires each group or researcher to construct a new database and therefore findings cannot be easily translated from one project to another, and the performance of different systems cannot readily be compared. A comprehensive review of this issue can be found in a paper by Cowie *et al*. (2001). Datasets of affective states differ in their scope (speakers and affective states), naturalness (acted, natural-like and natural expressions), context and annotation, the types of existing annotations for the database, such as expression labels and phonological descriptors (Douglas-Cowie *et al.* 2003). The development of databases to enhance the dynamic analysis of expressions should consider the same criteria.

The choice of datasets sets limits to the capabilities and scope of the developed system. Unfortunately, this choice is often dictated by the available resources. Many projects are based on staged expressions or on read paragraphs, using actors (Petrushin *et al.*, 1999). Another approach is to collect emotional episodes from films (Polzin *et al.*, 2000). Several speech databases include nonsense speech, with the aim of eliminating the effect of text. Most of these databases focus on Ekman's 'basic emotions' (Ekman, 1999), or on other small sets of extreme emotions. The problem is that extreme emotions are rare in everyday life, whereas nuances are common. Everyday expressions may include a mixture of intentions, mental states and emotions. In addition, staged expressions are different from real expressions; an example is the difference between a facial expression of smile and the label of happiness.

Several approaches for eliciting natural or natural-like emotions have been developed. One method is to use photographs, film episodes or music to elicit certain emotions. This method is used mainly for the recording of facial expressions. Another method is to record people who perform a given task, for example a frustrating computer game (Klein *et al.*, 2002), solving mathematical problems while driving (Fernandez & Picard, 2003) (for stress investigation) or asking young mothers to perform a certain task with their babies (Moore *et al.*, 1994). This method provides the researchers with more control on the

content and the setting in comparison to recordings of free speech, although most of the databases include only one modality, elicit a small variety of expressions, consider time-discrete events, and are proprietary.

There is a growing effort to use real recorded data for recognition (Vidrascu & Devillers, 2007, Xiao, 2007, Hoque & Louwerse, 2006.). The method that obtains data which is most natural is to use recordings of people in real situations, for example during telephone conversations, or of pilots during flight. The CREST database (Douglas-Cowie *et al.* 2003) is a major effort of this type and includes recordings of people in their natural environment over a period of five years. However, using corpora of real (not acted) recorded data for training often limits the scope of the system because annotation of real data is complicated (Douglas-Cowie *et al.* 2003, Devillers & Vidrascu & Lamel, 2005), which in practice limits the developers to labeling few affective states (or dimensions).

A common problem to all these methods is the association of names, labels, or descriptors with the recorded expressions. Often many subtle affective states may be defined under one definition of an affective state (Wierzbicka, 2000), for example different types of anger. Mixtures of affective states and co-occurring affective states also pose a problem for labeling. The dimensional approach, (Cowie & Cornelius, 2003) relates to descriptors of emotions. Tools such as FEELTRACE address labeling of emotional transitions along two axes, such as active-passive and positive-negative. These are only two of the multiple categories by which affective states can be defined. More elaborate systems, multi-modal analysis and context awareness can help the designers of a dataset to define additional descriptors. The W3C markup language which has recently been issued offers yet a wider range of descriptors (Burkhardt & Schröder, 2008). Several systems that should facilitate manual labeling of different aspects of multi-modal systems, such as video and speech, have been developed, but automatic systems are not available yet.

Another issue that arises from continuous recording is segmentation, i.e. dividing the speech into meaningful and manageable units. Further processing may involve recognizing different speakers and other related issues such as realizing when the speech is aimed at different listeners. 'Simple' segmentation, i.e. recognizing speech parts, has by itself been the focus of many research projects.

This paper presents two complementing datasets, one acted and labeled and the other comprising unlabelled naturally-evoked affective states, whose combination was used as part of a framework of an affect recognition system (Sobol-Shikler, 2010). This comprised the definition and extraction of vocal and temporal features of affective speech, design, training and testing of a system which infers the level of co-occurring affective states and for generalization of the system to a very wide variety of affective states and to the analysis of affect in sustained interactions in a different language.

2. Corpora

In order to achieve the defined goals of affect recognition, two complementing datasets, Mind Reading and Doors, were used.

Mind Reading

The Mind Reading database is classified using a prototypical taxonomy, and is available commercially as a DVD (Baron-Cohen *et al.*, 2004). This can be used to teach children and adults diagnosed with Autism Spectrum Disorder, who have difficulties recognizing emotional expression in others, to recognize the behavioral cues of a large variety of affective states in their daily lives. The dataset comprises three types of data: voice, video recordings of facial expressions and head movements and video recordings of body language and gestures in dialogues and in group interactions. An experimental version of the voice part of the database that consists of over 700 affective states was used. The database is arranged into 24 meaning groups, according to the Mind Reading taxonomy (Baron-Cohen *et al.*, 2004). The groups are: *afraid*, *touched*, *bothered*, *unfriendly*, *thinking*, *surprised*, *fond*, *hurt*, *sneaky*, *interested*, *angry*, *liked*, *sorry*, *bored*, *excited*, *sad*, *kind*, *disbelieving*, *wanting*, *sure*, *happy*, *romantic*, *unsure* and *disgusted*. Each of these groups comprises many affective states that share a meaning. Each affective state is represented by six different sentences uttered by six different actors with different (neutral) textual contents. In total, the dataset includes 4400 utterances recorded by ten UK English speakers of both genders and of different age groups, including children. According to its publishers, the acting was induced (Baron-Cohen *et al.*, 2004) and the database was labeled by ten different people. (The commercial version includes 412 of these affective states.)

The database is acted, but its original purpose (teaching humans) and the large number of affective states that it represents, make it a suitable choice for training a machine to recognize affective states and for validation on a large variety of affective states (although humans need fewer samples for training).

Doors

Doors (Sobol-Shikler, 2008) is a Hebrew database which was defined and recorded as part of this research in collaboration with the Psychology and the Bio-Engineering Departments in Tel Aviv University, in order to provide data and tools for different research projects. Doors is a multi-modal database of recorded sustained human-computer interactions. Doors comprises recordings of naturally evoked affective states in a controlled environment. The goal was to investigate affective non-verbal speech and facial expressions.

The participants were engaged in a computer game designed to evoke emotions, based on the Iowa Gambling Test (IGT) (Bechara *et al.*, 1997). Each interaction lasted approximately 15 minutes. The game comprised a series of 100 door opening events.

The speech part consists of two repeated sentences (/*ptah*

de'let zo/, open this door, and /*sgor de'let*/, close door, in Hebrew), forming for each participant a corpus of 200 sentences associated with the game events. After 20 trials, the participants were asked about their chosen strategy. This provided short intervals of speech towards people, with no text constraints. In addition, a few of the participants spoke freely during the interaction. The utterances with un-controlled text (comprising also sounds such as laughter and sighs) that were freely evoked during the game and during the intervening interviews varied in their number and nature among participants.

In addition to speech, the database comprises video recordings of facial expressions and head gestures, game events, (including participants' choices), mouse movement rate, reaction delay between events and physiological measurements, including: galvanic skin response (GSR), echo-cardiograms (ECG) and blood-volume at the periphery (BVP). These were recorded in order to support the identification and the labeling of affective states and for multi-modal analysis. Figure 1 represents a schematic description of the recording setup of the Doors database.

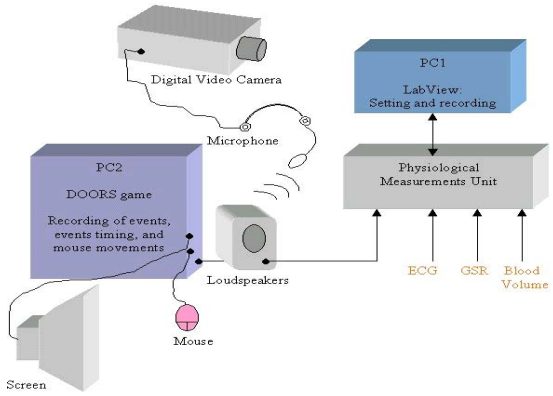

Figure 1 The recording setup of the Doors database.

The participants were Hebrew speaking graduate students and academic staff, of both genders in the age range of 24-55, mostly from engineering background, whose Hebrew is the first or second language.

3. Corpora in the Framework

The two datasets were used in different manners throughout the entire design process (Sobol-Shikler, 2010).

Feature Definition and Extraction

The first stage of the implementation entailed investigation of the vocal and temporal features of affective speech and developing reliable algorithms for their extraction. Both datasets were used in this stage. Doors provided a clear distinction of features relating only to affect, independent of the text. In particular, it showed subtle expressions in the context of HCI. It also provided the cues of dynamic changes between consecutive utterances during sustained interactions.

The Mind Reading dataset provided cues of a very large variety of affective states. This helped to define parameters that were not evident in the subtle expressions in Doors and set limits and extreme values of features, as well as extending the variety of speakers. Observations on both datasets revealed features and metrics, and the algorithms developed were tested on the two datasets. Normalization of each feature for each speaker was used in order to eliminate individual differences due to spoken language, accent, gender, and body structure. A set of 173 metrics of the defined features was identified and used for the classification (Sobol-Shikler, 2009).

Classification

The Mind Reading database was chosen for training and testing of the inference system. The database is acted, but its original purpose (teaching children to recognize affective states from their expressions in the children's daily lives) and the large number of affective states that it represents, make it a suitable choice for training a machine to recognize affective states and for validation on a large variety of affective states (although children need fewer samples for training). The database is labeled and provides a ground truth for the training and testing procedure.

The structure of the Mind Reading, which includes a large variety of concepts for each meaning group made the training set more similar to the variations of natural speech and transferrable between languages in which similar meaning groups exist (rather than individual concepts). Thus, the system was trained to infer affective-state groups (Sobol-Shikler & Robinson, 2010). Observations on the Doors database confirmed that several affective states can occur simultaneously and change asynchronously in time (for example: thinking and uncertainty, laughter with and without stress) (Sobol-Shikler, 2009). Therefore, the machine was trained to infer the level of expression of nine co-occurring affective-state groups. The chosen affective states are: *concentration*, *confidence*, *disagreement*, *excitement*, *interest*, *joy*, *stress*, *thinking* and *uncertainty*. They are common in everyday scenarios and were recognized in human computer interactions. Each group comprises of several affective states that share a meaning, as in the Mind Reading taxonomy, with modifications. Using groups of affective states had several advantages: It provided a wider definition of the inferred affective state; it compensated for the variability between speakers and languages and for the effects of acting, and it extended the number of samples used for training and testing while maintaining the inferred meaning.

The observations from the Doors database also provided the evidence that different sets of vocal cues distinguish the expressions of different pairs of affective states, and therefore there is no need to find a single subset of features to distinguish between all the affective states (Sobol-Shikler & Robinson, 2004). This observation was used to define the classification algorithm which used different sets of features to distinguish between different

pairs of affective states (36 classifications). Voting methods were applied to the combination of these classifications with an overall accuracy of over 83% (Sobol-Shikler & Robinson, 2010). An example of inference results for one sentence with the affective state *choosing* appears in Table 1.

Concentration	Confidence	Disagreement	Excitement	Interest	Joy	Stress	Thinking	Uncertainty
5	2	4	3	3	2	3	**7**	**7**

Table 1: Inference results of one sentence with the affective state "Choosing". The inferred affective-state groups (the voting results) are marked in grey.

Generalization

One of the research goals was to check if the recognition system could be generalized to a wide variety of affective states and to the analysis of sustained interactions.

A Large Variety of Affective States

The system was tested on the entire Mind Reading dataset (Sobol-Shikler & Robinson, 2010). For better accuracy, the accumulated results of all six sentences that represent each affective state were examined. Statistically significant consistency for all six sentences was found in 360 of affective states. Consistent results were found in at least four of the sentences (over one standard deviation above the mean number of sentences) for over 500 affective states. In 85% the inferred combination was also comparable to their lexical definitions and to the expected behavior associated with them. For example, the accumulated results of the sentences that represent the affective state *choosing* were a combination of *thinking* and *uncertainty*. The analysis of these results revealed interesting connections between meaning and affective behavior. It also allowed mapping of affective states according to the chosen affective-state groups. This provided a tool for verifying taxonomic representations of affect (Sobol-Shikler, 2009).

Multi-modal & Sustained Interactions

The system was also applied to the entire Doors database (Sobol-Shikler, 2008). Doors is mostly an unlabeled database. For the initial observations, utterances were manually annotated by several annotators, but the majority of the utterances comprised subtle affective states and nuances whose labeling was challenging. The additional measures of control in the recordings of the Doors database (same text, speaker, time, environment and equipment), as well as the multi-modal information, provided means to statistically analyze and validate the inference results. Each interaction was treated as a different test in which the recorded game events were correlated with the inferred affective state groups. Further temporal analysis was done in comparison to other recorded cues.

The inference system was applied automatically to each interaction (after segmentation into sentences and single utterances), with no discrimination between the utterances with the controlled and un-controlled text. No additional training of the inference machine was required, although the training was done on an English database while the Doors database was recorded in Hebrew. The only adjustment was the normalization of each vocal metric per speaker, as in the Mind Reading.

The IGT is a well established psychological test. Nevertheless, the participants of the Doors game who intentionally tried to find a strategy to increase their gain did not always react according to the reported results (Bechara *et al.*, 1997) and therefore could not be compared to them, nor could we identify a common behavior to all participants. The differences between the IGT game and the Door are the lack of 2000$ as a reward in the Doors game and the use of computer instead of cards. For example, one participant devised a strategy in which three of the doors were repeatedly chosen with impressive results until the 100th trial in which he lost everything. The strategy usually explored in the literature involves only two of the doors. The lack of high reward meant that the interaction more-closely resembled most human-computer interactions. This also meant that boredom and subtle affective states were prevalent. In addition, the goal was not to check the response to the Doors game but rather different manners of response in comparison to events and to other cues. Therefore, the results were compared within participant rather than between participants.

Because each interaction was a controlled experiment by itself, the entire inferred ranked list was used for the analysis, i.e. for each of the nine inferred affective-state groups (*concentration*, *confidence*, *disagreement*, *excitement*, *interest*, *joy*, *stress*, *thinking* and *uncertainty*), was assigned the level of recognized expression in the range 0-8.

The statistical analysis was divided into different questions, such as analyzing each of the two repeated sentences in relation to the corresponding events, and analyzing all the utterances before and after certain events. Examples of analyzed events include: temporal gain (positive or negative), total gain above and below zero, participants' choice of advantageous and disadvantageous doors, before and after the interview, and the like. The results for each participant were different, but all showed significant influence and corresponding vocal and affective responses to the events of the game, and accompanied to their own choices (Sobol-Shikler, 2008, Sobol-Shikler, 2009).

After the initial statistical analysis that proved significant relations between the inference results and the interaction events, temporal analysis was performed. This involved the analysis of tendencies of the inferred affective states over time and their co-occurrences with events such as gain changes over time and changes in physiological cues and other behavioral cues such as mouse movements and

reaction delays. The last stage was finding temporal events and their corresponding vocal reaction. This is limited to a few significant events, for example the relation between the verbal content of utterances with un-controlled text and the affect inferred from their vocal cues (Sobol-Shikler, 2008). These cannot be justified in a statistical manner, but their accuracy is interesting and further suggests that such analysis is justified and should be extended.

4. Conclusions

The paper presented a framework which comprised all stages of design and validation of a recognition system of affect from non-verbal speech, including the definition of prosodic cues of affect, implementation of a system which infers the level of co-occurring affective states, and analysis of the relations between the recognized affective states in meaning and in time. The paper showed how the implementation of this framework was enabled by the choice and definition of two complementing datasets, Mind Reading and Doors.

Mind Reading provides labeled samples of a very large variety of complex affective state concept, arranged into meaning groups, by speakers of different age groups. This enabled classification of affective-state groups, and also enabled testing inference of co-occurring affective states.

Doors comprised affective states and dynamic changes of affect naturally-evoked during sustained human-computer interaction, controlled text and un-controlled text, and multi-modal information.

The two datasets provided over 7000 sentences by different speakers in two different languages, including 4400 acted and labeled sentences in English, over 2700 text repetitions and around 100 utterances with un-controlled text of naturally evoked affective states in Hebrew, uttered by 25 speakers of both genders and different age groups, actors and non-actors.

This combination of datasets contributed throughout the development of the inference system, contributed to its generality and robustness and revealed important features of affective behavior and its meaning, extending the scope of the system toward universality. In addition, it showed that the vocal cues and behavior related to complex affective states transcend language barriers.

5. Acknowledgement

The author thanks Rinat Bar-Lev, Matti Mintz, Shay Davidi and Oded Barnea for their part in the recording of the Doors database. The author thanks AAUW Educational Foundation, Cambridge Overseas Trust and Deutsche Telekom Laboratories for their partial support of this research.

6. References

Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.J. (2004). *Mindreading: The interactive guide to emotions*. London, UK: Jessica Kingsley Limited. (http://www.jkp.com).

Bechara, A., Damasio, H., Tranel, D., Damasio, A.R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275, pp. 1293--1295.

Burkhardt, F., Schröder, M. (2008). Emotion markup language: Requirements with priorities. *W3C Incubator Group*. (http://www.w3.org/2005/Incubator/emotion.)

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18, pp. 32--80.

Cornelius, R., Cowie, R. (2003). Describing the Emotional States that are Expressed in Speech. *Speech Communication*, 59.

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., Campbell, N., Cowie, R., Roach, P. (2003). Emotional speech: towards a new generation of databases. *Speech Communication*, 40, pp. 33--60.

Ekman, P. (1999). Basic emotion. In M., Power, & T., Dalgleish, (Eds.), *Handbook of cognition and emotion*. Chihester, UK: Wiley.

Fernandez, R., Picard, R. W. (2003). Modeling drivers' speech under stress. *Speech Communication*, 40, pp. 145--159.

Hoque, M. Y. M., Louwerse, M. (2006). Robust recognition of emotion from speech. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, Marina del Rey.

Klein J., Moon Y., Picard R. W. (2002). This computer responds to user frustration: theory, design, and results. *Interacting with Computers*, 14(2), pp. 119--40.

Moore, C. A., Cohn, J. F., Katz, G. S. (1994). Quantitative description and differentiation of fundamental frequency contours. *Computer Speech & Language*, 8(4), pp. 385--404.

Petrushin V. (1999). Emotion in speech: Recognition and application to call centers. *Intelligent Engineering Systems Through Artificial Neural Networks*, pp. 1085--1092.

Picard, R.W. (1997). *Affective Computing*. Boston, MA: MIT Press.

Polzin, T., Waibel A. (2000). Emotion-sensitive human-computer interfaces. In *Proceedings of the ISCA workshop on speech and emotion*. Belfast, North Ireland.

Sobol-Shikler, T., Robinson, P. (2004). Visualizing Dynamic Features of Expressions in Speech. In *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju, Korea.

Sobol-Shikler, T. (2008). Multi-modal analysis of human computer interaction using automatic inference of aural expressions in speech. In *Proceedings of IEEE Systems Man & Cybernetics*, Singapore.

Sobol-Shikler, T. (2009). Analysis of affective expressions in speech. Technical Report. University of Cambridge.

Sobol-Shikler, T., Robinson, P. (2010). Classification of complex information: Inference of co-occurring

affective states from their expressions in speech. In press in *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 32. (10.1109/TPAMI.2009.107).

Sobol-Shikler, T. (2010). Automatic Inference of Complex Affective States. In press in *Computer, Speech and Language*. (10.1016/j.csl.2009.12.005).

Vidrascu, L., Devillers, L. (2007). Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features. In *Proceedings of the International workshop on Paralinguistic Speech*, Saarbrcken, Germany, pp. 11--16.

Wierzbicka, A. (2000). The semantics of human facial expressions. *Pragmatics & Cognition*, 8(1).

Xiao, Z., Dellandrea, E. , Dou, W., Chen, L. (2007). Automatic hierarchical classification of emotional speech. In *Proceedings of the Multimedia Workshops, ISMW'07*, pp. 291--296.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 31, pp. 39--58.

Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments

Florian Eyben[1], Anton Batliner[2], Björn Schuller[1], Dino Seppi[3], Stefan Steidl[2]

[1]Institute for Human-Machine Communication, Technische Universität München,
[2]Pattern Recognition Lab, FAU Erlangen, [3]ESAT, Katholieke Universiteit Leuven,
[1]München, Germany, [2]Erlangen, Germany, [3]Leuven, Belgium
eyben@tum.de, batliner@informatik.uni-erlangen.de

Abstract

We use four speech databases with realistic, non-prompted emotions, and a large state-of-the-art acoustic feature vector, for cross-corpus classifications, in turn employing three databases for training and the fourth for testing. Categorical and continuous (dimensional) annotation is mapped onto a representation of valence with the three classes positive, neutral, and negative. This cross-corpus classification is compared with within corpus classifications. We interpret performance and most important features.

1. Introduction

The normal approach towards classifying emotion in speech is to subdivide one corpus into specific train, validation and test subsets, in the case of cross-validation with or without using specific validation sets. By that, many intervening variables such as microphone, room acoustics, speaker group, etc., are kept constant. However, we always have to keep in mind that we rather cannot generalize onto other corpora and settings when using this approach. A first step towards overcoming such restrictions and thus evaluating recognition of realistic emotions in a scenario which itself is more realistic, is doing cross-corpus classification. This will be pursued in this paper. First, in section 2., we introduce the four naturalistic emotion corpora used in this study. In section 3., we describe our acoustic feature set, and in section 4., we present results and describe the evaluation methods. Concluding remarks are given in section 5.

2. Corpora

Table 1 shows the basic statistics of the four naturalistic emotion corpora used in this study, namely the SmartKom Corpus (SmK), the FAU Aibo Emotion Corpus (Aibo), the Sensitive Artificial Listener Corpus (SAL), and the Vera-am-Mittag Corpus (VAM). One of the main difficulties of cross-corpus experiments in this field, besides the different content and acoustics, is the mismatch of annotations with respect to the labels considered. Each corpus was recorded more or less for a specific task – and as a result of this, they have specific emotion labels assigned to them. For cross-corpus recognition this poses a problem, since the training and test sets in any classification experiment must use the same class labels. This is especially problematic for corpora where annotations are made in terms of discrete class labels, such as SmartKom and Aibo. Corpora annotated in terms of affect dimensions such as valence and arousal are easier to match, although per corpus biases and different ranges can pose a problem.

In order to be able to perform cross-corpus valence recognition in this study, a standard set of classes has been defined for all corpora; we decided for three levels of valence: negative, idle (i. e. neutral), and positive. The labels used in each corpus can be mapped onto these three classes. This mapping can be found in the following subsections for each corpus. Moreover, a description of the notion of 'turn', which is used as unit of analysis, can be found in the corpus documentation in the following subsections. Table 1 reveals that there is indeed a considerable variation in turn duration and by that, most likely in consistency of valence.

2.1. SmartKom

SmartKom (SmK) is a multi-modal German dialogue system which combines speech with gesture and facial expression. The so called SmartKom-Public version of the system is a 'next generation' multi-modal communication telephone booth. The users can get information on specific points of interest, as, e.g., hotels, restaurants, cinemas. They delegate a task, for instance, finding a film, a cinema, and reserving the tickets, to a virtual agent which is visible on the graphical display. Users get the necessary information via synthesised speech produced by the agent, and on the graphical display, via presentations of lists of points of interest (e. g. hotels, restaurants, and cinemas), and maps of the inner city. For this system, data are collected in a large-scaled Wizard-of-Oz (WoZ) experiment. The dialogue between the (pretended) SmK system and the user is recorded with several microphones and digital cameras. Subsequently, several annotations are carried out. The recorded speech represents thus a special variety of non-prompted, spontaneous speech typical for man-machine-communication in general and for such a multi-modal setting in particular. More details on the recordings and annotations can be found in (Steininger et al., 2002; Batliner et al., 2003). The labellers could look at the persons' facial expressions, body gestures, and listen to his/her speech; they annotated the user states *joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise, neutral*, and *unidentifiable* episodes. *Joy* and *anger* were subdivided into the subclasses *weak* and *strong joy/anger*. The labelling was frame-based, i. e. beginning and end of an emotional episode was marked on the time axis. Turns are defined as dialogue moves, i. e. as everything produced by the user until the system takes over.

We mapped the class *anger* to negative valence (**N**), the classes *helplessness, pondering*, and *neutral* to neutral va-

Corpus	# of instances				Turn duration (s)			
	P	**I**	**N**	Overall	Mean	Stddev.	Min	Max
SmK	353	2963	219	3535	6.8	7.1	0.1	64.2
Aibo	495	11021	2215	13731	2.3	1.5	0.9	38.0
SAL	466	588	638	1692	3.5	3.0	0.9	26.8
VAM	16	511	420	947	3.0	2.2	0.4	17.7

Table 1: Number of instances in each of the four corpora; distribution of instances among the 3 valence classes (**N**: negative valence, **I**: idle, i. e. neutral valence, **P**: positive valence), and mean, minimum, and maximum turn duration per corpus.

lence (**I**), and *joy* and *surprise* to positive valence (**P**). The unidentifiable episodes were ignored, since they might contain episodes with positive or negative valence which could not be mapped onto the pre-defined classes.

2.2. Aibo

The FAU Aibo Emotion Corpus comprises recordings of German children's interactions with Sony's pet robot Aibo; the speech data are spontaneous and emotionally coloured. The children were led to believe that the Aibo was responding to their commands, whereas the robot was actually controlled by a human operator. This WoZ caused the Aibo to perform a fixed, predetermined sequence of actions; sometimes the Aibo behaved disobediently, thereby provoking emotional reactions. The data was collected at two different schools, MONT and OHM, from 51 children (age 10 - 13, 21 male, 30 female; about 8.9 hours of speech without pauses > 1 s). Speech was transmitted with a high quality wireless head set and recorded with a DAT-recorder (16 bit, 48 kHz down-sampled to 16 kHz). The recordings were segmented automatically into 'turns' using a pause threshold of 1 s. Five labellers listened to the turns in sequential order and annotated each word independently from each other as neutral (default) or as belonging to one of ten other classes. We resort to majority voting (MV): if three or more labellers agreed, the label was attributed to the word. In the following, the number of cases with MV is given in parentheses: *joyful* (101), *surprised* (0), *emphatic* (2 528), *helpless* (3), *touchy*, i. e. irritated (225), *angry* (84), *motherese* (1 260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral, but not belonging to the other categories (3), *neutral* (39 169); 4 707 words had no MV; all in all, there were 48 401 words. *reprimanding*, *touchy*, and *angry* were mapped onto a main class *angry*. The mapping of word- onto turn-labels is described in (Steidl, 2009).

We map (based on the turn labels) the classes *angry* and emphatic to negative valence (**N**), the classes *neutral and rest* to neutral valence (**I**), and *motherese and joyful* to positive valence (**P**). *Helpless*, *surprised*, and *bored* did not occur amongst the turn based labels.

2.3. SAL

The Belfast Sensitive Artificial Listener (SAL) data is part of the final HUMAINE database (Douglas-Cowie et al., 2007). We consider the subset used e. g. in (Wöllmer et al., 2008) which contains 25 recordings in total from four speakers (2 male, 2 female) with an average length of 20 minutes per speaker. The data contains audio-visual recordings from human-computer conversations (WoZ scenario) that were recorded through a SAL interface designed to let users work through a range of emotional states. The data has been labelled continuously in real time by four annotators with respect to valence and activation using a system based on FEELtrace (Cowie et al., 2000): the annotators used a sliding controller to annotate both emotional dimensions separately whereas the adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. To compensate linear offsets that are present among the annotators, the annotations were normalised to zero mean globally. Further, to ensure common scaling among all annotators, each annotator's labels were scaled so that 98 % of all values are in the range from –1 to +1. The 25 recordings have been split into turns using an energy based Voice Activity Detection. Accordingly, a total of 1 692 turns is contained in the database. Labels for each turn are computed by averaging the frame level valence and activation labels over the complete turn.

We define the classes negative valence (**N**) for turns with an annotated valence below -0.25, neutral valence (**I**) from -0.25 to 0.25, and positive valence (**P**) for turns with an annotated valence above 0.25.

2.4. VAM

The Vera-Am-Mittag (VAM) corpus (Grimm et al., 2008) consists of audiovisual recordings taken from a German TV talk show. The set used contains 947 spontaneous and emotionally coloured turns from 47 guests of the talk show which were recorded from unscripted discussions. The topics were mainly personal issues such as friendship crises, fatherhood questions, or romantic affairs. To obtain non-acted data, a talk show in which the guests were not being paid to perform as actors was chosen. The speech extracted from the dialogues contains a large amount of colloquial expressions as well as non-linguistic vocalisations and partly covers different German dialects. For annotation of the speech data, the audio recordings were manually segmented into turns, each utterance containing at least one phrase. A large number of human labellers was used for annotation (17 labellers for one half of the data, six for the other). The labelling bases on a discrete five point scale for three dimensions mapped onto the interval of [–1,1]: the average results for the standard deviation are 0.29, 0.34, and 0.31 for valence, activation, and dominance. The averages for the correlation between the evaluators are 0.49, 0.72, and 0.61, respectively. The correlation coefficients for activation and dominance show suitable values, whereas the moderate value for valence indicates that this emotion primitive was more difficult to evaluate; it may partly also

be a result of the smaller variance of valence.
As for the SAL corpus, we define the classes negative valence (**N**) for turns with an annotated valence below -0.25, neutral valence (**I**) from -0.25 to 0.25, and positive valence (**P**) for turns with an annotated valence above 0.25.

2.5. Inconsistencies across corpora

When doing cross-corpus classification using these four corpora, we are facing several problems and inconsistencies that most certainly will not always be favourable for our classification performance:

- Three of the corpora are German, one, i. e. SAL, is English.
- In three corpora, speakers are adult, whereas in Aibo, speakers are children.
- The scenarios differ in several respect: SmK is about information queries in a human-WoZ setting; Aibo is about giving commands to a pet robot, again in a WoZ setting – however, the WoZ (Aibo) never talks; SAL is about the interaction with an emotional agent, again in a WoZ setting; VAM is about human-human interaction in an 'emotion-prone' talk show.
- Subjects are 'naive' in SmK and in Aibo, experts in SAL, and most likely belonging to some specific type of personality in VAM.
- Number of subjects, of labellers, and of items per class can differ considerably.
- The original units of annotation differ: frames in SmK, words in Aibo, turns in SAL and in VAM; different types of mappings onto the turn level had to be performed. Certainly, this goes along with less clear, 'smeared' classes, although this might have different impact in each of our four corpora.
- The emotional taxonomies differ: categories in SmK and in Aibo, dimensions in SAL and in VAM; again, we had to perform different types of mapping onto our three mutually exclusive valence classes. As a consequence, a few 'garbage' turns had to be mapped or skipped on a somehow arbitrary basis.
- Last but not least, valence – both as dimension or as categories – is notoriously more difficult to process and classify than, e. g. arousal, when only acoustic information is used, because a straightforward equation such as 'higher/longer/stronger means higher arousal, and vice versa' cannot be used.

3. Acoustic features

We use a set of 2 832 acoustic features extracted with the openEAR toolkit (Eyben et al., 2009). Thereby 59 acoustic low level descriptor (LLD) contours (see table 2) are computed at a rate of one every 10 ms. A Gaussian window ($\sigma = 0.25$) of size 50 ms is used for all LLD except for pitch and formants, where a window size of 75 ms is preferred. A pre-emphasis with a factor of $k = 0.97$ is applied to the 50 ms frames, and a de-emphasis with factor $k = 0.92$ is applied to the 75 ms frames.
First order delta regression coefficients are computed from all 59 LLD contours resulting in 118 LLD features in total. After applying the 24 functionals described in table 3 to each of the LLD, a 2 832 dimensional vector is obtained for each input instance (turn).

Feature Group	**Features in Group**
Pitch	F_0 in Hz via sub-harmonic sampling (F_0), smoothed F_0 contour (Hz) ($F_0 env$)
Energy	Intensity ($Intens$)
Formant	Formant frequency ($freq$) and bandwidth (bw) of F_1 to F_4 via LPC analysis, LPC gain
Voice Quality	Probability of voicing (p_{voice}), local Jitter (Jit_{loc}), differential Jitter (Jit_D), local Shimmer (Shi_{loc})
Spectral	Centroid, Entropy, Flux 90 % roll-off point (rop) and position of highest peak in spectrum ($specMaxPos$).
Mel-bands	Mel-frequency-bands (MFB) 0-25 (20-8000 Hz)
Cepstral	MFCC 1–12

Table 2: Set of 59 Low-Level Descriptors (LLD).

Functionals	**Abbrv.**
Maximum and Minimum value	max/min
Range (Max.–Min.)	range
Arithmetic Mean (of non-0 values)	(nz)amean
Relative pos. of global max. value	maxpos
Centroid	centroid
Linear regression coefficients and corresp. quad. approximation error	qregc1–3 qregerr
Quadratic regression coefficients and corresp. quad. approximation error	linregc1–2 linregerr
Number of non-zero values	nnz
Standard deviation	stddev
Skewness, kurtosis	skew/kurt
Number of peaks	numPeaks
Arithmetic mean of peaks	peakMean
Mean distance between peaks	meanPeakDist
Rel. time below 25% of range	downleveltime25
Rel. time above 75%/90% of range	upleveltime75/90

Table 3: Set of 24 functionals applied to LLD contours and delta coefficients of LLD contours. Abbreviations as used in the following tables.

4. Classification and Results

In total we perform three experiments: within-corpus classification, cross-corpus classification (leave one corpus

[UAR %]	PI-N	P-IN	PN-I	P-N	Avg.
SmK	50.7*	49.5	56.3*	58.9	53.9
Aibo	57.9*	57.6*	59.9*	76.0*	62.9*
SAL	61.9	53.1	46.5	69.5*	57.8
VAM	60.2*	(50.0)	58.1	(50.0)	(54.6)
Avg.	57.8*	52.5	55.2	63.6	57.3

Table 4: Within corpus UAR obtained on four corpora with SVM (SMO). * indicates significant improvement ($\alpha = 0.01$) over random guess. 10-fold SCV.

[UAR %]	PI-N	P-IN	PN-I	P-N	Avg.
SmK	51.1*	52.4	47.9	55.0*	51.6
Aibo	52.0	55.7	50.7	54.9*	53.3
SAL	52.2*	49.0	51.9	48.6	50.4
VAM	56.2*	63.4*	53.8*	59.1*	58.1*
Avg.	52.9	55.1	51.1	54.4*	53.4

Table 5: cross-corpus UAR obtained on four corpora as test sets (leave-one-corpus-out) with SVM (SMO). * indicates significant improvement ($\alpha = 0.01$) over random guess.

out), and cross-corpus feature ranking. For establishing a coarse, preliminary reference for within corpus classification, we perform 10-fold cross validation (note that this is not speaker independent). We use Support Vector Machines (SVM) trained with the Sequential Minimal Optimisation algorithm (SMO) as implemented in WEKA 3 (Witten and Frank, 2005) for all experiments. In order to obtain a somewhat generalised and classifier independent feature ranking, we use Discriminative Multinominal Bayes (DMNB) (Su et al., 2008) and Support Vector Machines as implemented by LibSVM (Chang and Lin, 2001) in addition to the SMO SVM.

To investigate the performance for classifying different aspects of valence independently, we map the three classes **P**, **I**, and **N** onto four binary class sets: **P** and **I** vs. **N** (**PI-N**), **P** vs. **N** (**P-N**), **P** vs. **I** and **N** (**P-IN**), and **P** and **N** vs. **I** (**PN-I**). Doing that, on the one hand we subdivide the valence axis at two different points: between positive and rest (**I** and **N**), and between negative and rest (**I** and **P**). On the other hand we know that neutral (**I**) is very often confused with either positive or negative, cf. (Batliner et al., 2008), thus we evaluate the performance of contrasting **P** vs. **N** leaving aside **I**, and telling apart **I** from emotional (**P** and **N**).

4.1. Within corpus evaluation

For each of these four sets we compute within corpus recognition results in terms of unweighted average class-wise recall rate (UAR) by 10-fold stratified cross validation (SCV); UAR is computed as the mean value of the numbers of correctly recognised instances per class divided by the total number of instances per class. By that the resulting numbers are not biased by the distribution of instances among classes. These results are given in table 4. For this preliminary within corpus classification, we decided not to balance the number of instances for this experiment, since for VAM only 16 **P** instances exist and thus balancing via sub-sampling is not feasible. This, however, yields non-informative results for the sets **P-IN** and **P-N** on the VAM corpus (last line of table 4). Leaving aside VAM, the quality of the performance is positively correlated with the turn length (cf. table 1): the shorter the turns are, the more likely it is that they are emotionally consistent, i. e. that the emotion is constant throughout the turn. Note that due to the small number of corpora, this is no hard proof yet but an indication worthwhile to be pursued further.

4.2. Cross-corpus evaluation

Next, we report cross-corpus results in table 5. One of four corpora was used for testing while the other three were combined as training set (the sets are speaker disjunctive, thus the results indicate speaker and corpus independent performance). For this experiment, the training set is balanced by randomly sub-sampling all classes to the number of instances in the smallest class. The distribution of instances among classes in the test set, however, is not balanced. Thus, we prefer the unweighted average class-wise recall rate (UAR) as an evaluation metric.

In contrast to within-corpus classification, there is no clear-cut correlation between performance and emotional consistency. This could be expected because training and test set differ with respect to several factors as detailed in section 2.5. Moreover, the average length of turns differ within the training set and across training and test set.

4.3. Cross-corpus feature ranking

The two experiments described so far were conducted with the full set of 2 832 features. We now want to find generic, corpus independent acoustic features relevant for revealing valence in general, and for each of the four binary class sets in particular. Since an exhaustive search on a set of 2 832 features is not feasible in a decent amount of time with today's hardware, we perform a quick estimation of the impact of each low-level descriptor and each functional separately. For this we evaluate the classification performance (UAR) of 142 individual feature sets. 118 sets are created by extracting single LLD with all 24 functionals applied to them. The remaining 24 sets are created by applying each of the 24 functionals separately to all 118 LLD. We then rank the 118 and 24 sets by UAR and thus obtain two rankings, one for LLD and one for functionals. For each of the three classifiers we obtain a separate ranking, as well as for each of the four corpora. Thus we obtain $3 \cdot 4 = 12$ rankings of LLD and functionals for each binary class set. We then compute the mean rank of each feature over all 12 rankings to obtain a unified ranking for each binary class set. The mean rank over all four class sets gives the overall rank of features for valence recognition. For the final sets of relevant features we select only those which by themselves achieve an UAR performance of ≥ 0.51. Table 6 shows the top 5 selected functionals; in table 7 we show the top 10 selected LLD for the four class sets. This roughly amounts to $\frac{1}{5}$ of the 24 functionals and the 59 LLD.

Since this feature ranking is only uni-variate and features with similar rank may still be correlated, this contribution should be considered as a pilot study, and a more thorough

Set	Functionals	# sel.
All	upleveltime75, downleveltime25, amean, kurtosis, min, ...	18
PI-N	nzamean, min, amean, downleveltime25, upleveltime75, ...	17
P-IN	upleveltime75, upleveltime90, qregerr, range, qregc1	5
PN-I	min, max, numPeaks, linregerr, amean, ...	8
P-N	upleveltime75, downleveltime25, skewness, kurtosis, range, ...	22

Table 6: Top 5 of selected functionals, and number of selected functionals in total (individual UAR ≥ 0.51).

Set	LLD	# sel.
All	MFCC 1,3–5,8,10,12 MFB 19,20, spec. Flux	56
PI-N	MFCC 1–5,8,10–12, MFB 20, spec. Flux	33
P-IN	MFB 6,8–12,18,23-25	32
PN-I	MFCC 1,4,5,6,7,8,12 MFB 20,21, spec. Flux	22
P-N	P_{voice}, MFCC 1,3,4,10, MFB 14,19, $F_{2,3,freq.}$, spec. Flux	79

Table 7: Top 10 of selected LLD, and number of selected LLD in total (individual UAR ≥ 0.51).

search for the best feature set must be conducted in future work. However, from the rankings of the LLD and the functionals, a slight tendency across all class configurations can be observed. For the functionals up-/downlevel-times (esp. upleveltime75) prevail in the top 5 for all configurations except for **PN-I**. This configuration is quite different with respect to selected functionals. This seems logical, since **PN-I** is about discriminating positive/negative valence from neutral, while the other three configurations are about detecting positive or neutral valence vs. the rest. For the top 10 LLD, the picture is quite different. The **PN-I** configuration is not exceptional. Instead, for the **P-IN** configuration a high relevance of only MFB is observed. For the three other configurations, spectral flux (i. e. the spectral difference between consecutive frames), higher order MFB (above 19, or 20) and lower order MFCC, esp. 1, and 4, occur frequently in the top 10 list. The higher order MFB correspond to frequencies in the 4–6 kHz region, where the upper formants are found.

With respect to the number of selected LLD/functionals, the **P-N** configuration is leading, which is in line with the finding that **P-N** performs best in overall classification (second for cross-corpus and best for within corpus), when we consider the uni-variate selection process.

As expected, the within corpus results are better than the cross-corpus results, yet the difference is only approx. 4% on average. Within corpus recognition for the **P-IN** constellation is below cross-corpus performance. The biggest difference can be observed again for the **P-N** configuration. This is another indicator that the separation of the classes **P** and **N** is the most doable.

5. Discussion and concluding Remarks

We have presented pilot experiments in a novel field: cross-corpus recognition of naturalistic emotions (here: valence) from acoustic features. Significant improvements over random guess were observed in at least a few cases, which indicates that cross-corpus recognition – even, for acoustic feature based approaches, of the most challenging dimension valence – is feasible in principle; however, it needs more effort to mature to a usable stage. Separation of positive vs. neutral valence gave best results, while a neutral (idle) vs. rest scenario showed lower recall rates. This indicates a fundamental problem with naturalistic emotion recognition: emotions are a continuum. Tagging emotions with discrete classes works for prototypical emotions (such as **P** and **N** valence), but yields inherent quantisation errors when dealing with naturalistic emotions, where there is no fixed class border and thus confusions between adjacent classes are common.

It is generally known that valence recognition from acoustic cues alone is challenging and perhaps not possible perfectly. The within corpus recognition results support this, as well as other studies on the SAL and VAM databases (Wöllmer et al., 2008; Grimm et al., 2007). Thus, future studies might need to investigate linguistic features as well as other modalities, such as vision. Moreover, these studies need to consider word-level chunking, which also has been proven to yield better results (Batliner et al., 2010).

A necessary step towards improving classification performance will be to take care of the inconsistencies listed in section 2.5. Some of these inconsistencies are given (different languages/scenarios) or can only be minimized with a high effort, such as differences in type of labels and number of annotators. However, we can try and find out which corpora are 'good', and which are 'bad' to be included in such cross-corpus evaluations; in other words, which are generic enough, and which are too specific. And we can define the same and optimal type of unit of analysis, for instance words or syntactically well-defined chunks, across all corpora.

6. Acknowledgment

The research leading to these results has received funding from the European Community under grant (FP7/2007-2013) No. 211486 (SEMAINE), grant No. IST-2002-50742 (HUMAINE), and grant No. IST-2001-37599 (PF-STAR). The responsibility lies with the authors.

7. References

A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth. 2003. We are not amused - but how do you know? User states in a multi-modal dialogue system. In *Proc. Interspeech*, pages 733–736, Geneva.

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.

A. Batliner, D. Seppi, S. Steidl, and B. Schuller. 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human-Computer Interaction*, 2010. doi:10.1155/2010/782802.

C.-C. Chang and C.-J. Lin, 2001. *LibSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 2000. Feeltrace: An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24, Newcastle, Northern Ireland.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 488–500, Berlin-Heidelberg. Springer.

F. Eyben, M. Wöllmer, and B. Schuller. 2009. openear - introducing the munich open-source emotion and affect recognition toolkit. In *Proc. ACII*, pages 576–581, Amsterdam.

M. Grimm, K. Kroschel, and S. Narayanan. 2007. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Proc. ICASSP*, pages IV–1085–IV, Honolulu.

M. Grimm, Kristian Kroschel, and Shrikanth Narayanan. 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, Hannover, Germany.

S. Steidl. 2009. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin. (PhD thesis, FAU Erlangen-Nuremberg).

S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. 2002. Development of user-state conventions for the multimodal corpus in smartkom. In *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 33–37, Las Palmas.

J. Su, H. Zhang, C.X. Ling, and S. Matwin. 2008. Discriminative Parameter Learning for Bayesian Networks. In *Proc. ICML*, pages 1016–1023, Helsinki.

I. H. Witten and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.

M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. 2008. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. Interspeech*, pages 597–600, Brisbane.