The Workshop Programme

Corpora for Research on Emotion and Affect
Tuesday 23rd May 2006
14:30 – 20:00

14:30-14:50 Welcome (R. Cowie)

14:50-16:30 First Oral Session (15 min. per speaker + 5 min. for questions)

14:50 Nick Campbell
A Language-Resources Approach to Emotion: Corpora for the Analysis of Expressive Speech

15:10 Franck Enos, Julia Hirschberg
A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process

15:30 Magalie Ochs, Catherine Pelachaud, David Sadek
A Coding Scheme for Designing: Computational Model of Emotion Elicitation

15:50 Tanja Bänziger, Hannes Pirker, Klaus Scherer
GEMEP - GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions

16:10 Laurence Vidrascu, Laurence Devillers
Real-life emotions in naturalistic data recorded in a medical call center

16:30-17:00 Coffee break

17:00-18:00 Poster Session

Poster 1: Noam Amir, Samuel Ron
Collection and evaluation of an emotional speech corpus using event recollection

Poster 2: M. Rita Ciceri, Stephania Balzarotti, F. Manzoni
MEED: the challenge towards a Multidimensional Ecological Emotion Database

Poster 3: Steffi Frigo
The relationships between acted and naturalistic emotional corpora

Poster 4: Nicolas Audibert, DamienVincent, Véronique Aubergé, Olivier Rosec
Evaluation of expressive speech resynthesis

Poster 5: Isabella Poggi
Body and Music. An annotation scheme of the pianist multimodal behaviour

Poster 6: Aubergé Véronique, Rilliard Albert, Audibert Nicolas
Auto-annotation: an alternative method to label expressive corpora

Poster 7: Loyau Fanny, Aubergé Véronique
Expressions outside the talk turn: ethograms of the feeling of thinking

Poster 8: Valérie Maffiolo, G. Damnati, V. Botherel, E. Guimier de Neef and E. Maillebuau
Multilevel features for annotating application-driven spontaneous speech corpora

Poster 9 : Maartje Schreuder, Laura van Eerten, Dicky Gilbers
Music as a method of identifying emotional speech

Poster 10: Aaron S. Master, Ing-Marie Jonsson, Clifford Nass, Peter X. Deng, Kristin L. Richards
A Framework for Generating and Indexing Induced Emotional Voice Data

Poster 11: Alexander Osherenko
Affect Sensing using Lexical Means: Comparison of a Corpus with Movie Reviews and a Corpus with Natural Language Dialogues

Poster 12: N. Mana, P. Cosi, G. Tisato, F. Cavicchio, E. Magno and F. Pianesi
An Italian Database of Emotional Speech and Facial Expressions

18:00-19:40 Second Oral Session (15 min. per speaker + 5 min. for questions)

18:00 Vaishnevi Varadarajan, John Hansen, Ikeno Ayako
UT-SCOPE – A corpus for Speech under Cognitive/Physical task Stress and Emotion

18:20 Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, Thibaul Ehrette
The SAFE Corpus: illustrating extreme emotions in dynamic situation

18:40 Janne Bondi Johannessen, Kristin Hagen, Joel Priestley and Lars Nygaard
A speech corpus with emotions

19:00 Dirk Heylen, Dennis Reidsma, Roeland Ordelman
Annotating State of Mind in Meeting Data

19:20 Mark Schröder, Hannes Pirker, Myriam Lamolle
First suggestions for an emotion annotation and representation language

19:40-20:00 Conclusion and Panel/Discussion (20mn)

Workshop Organisers

Laurence Devillers / Jean-Claude Martin,
Spoken Language Processing group/ Architectures and Models for Interaction,
LIMSI-CNRS, France

Roddy Cowie / School of Psychology
Ellen Douglas-Cowie / Dean of Arts, Humanities and Social Sciences
Queen's University, Belfast BT7 1NN, UK

Anton Batliner - Lehrstuhl fuer Mustererkennung (Informatik 5)
Universitaet Erlangen-Nuernberg, Germany

Workshop Programme Committee

Elisabeth André, Univ. Augsburg, D
Véronique Aubergé, CNRS-STIC, FR
Anton Batliner, Univ. Erlangen, D
Nadia Bianchi-Berthouze, Univ. Aizu, J
Nick Campbell, ATR, J
Roddy Cowie, QUB, UK
Laurence Devillers, LIMSI-CNRS, FR
Ellen Douglas-Cowie, QUB, UK
John Hansen,Univ. of Texas at Dallas, USA
Susanne Kaiser, UNIGE, S
Stephanos Kollias, ICCS, G
Christine Lisetti, EURECOM, FR
Valérie Maffiolo, FranceTelecom, FR
Jean-Claude Martin, LIMSI-CNRS, FR
Shrikanth Narayanan, USC Viterbi School of Engineering, USA
Catherine Pelachaud, Univ. Paris VIII, FR
Isabella Poggi, Univ. Roma Tre, I
Fiorella de Rosis, Univ. Bari, I
Izhak Shafran, Univ. Johns Hopkins, CSLP, USA
Elisabeth Shriberg, SRI and ICSI, USA
Marc Schröder, DFKI Saarbrücken, D
Ioana Vasilescu, ENST, FR

Table of Contents

Author Index

Corpora for research on Emotion and Affect

Introduction

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

The HUMAINE network of excellence (http://emotion-research.net/) has brought together several groups working on the development of databases, and the workshop aims to broaden the interaction that has developed in that context.

Many models of emotion are common enough to affect the way teams go about collecting and describing emotion-related data. Some which are familiar and intuitively appealing are known to be problematic, either because they are theoretically dated or because they do not transfer to practical contexts. To evaluate the resources that are already available, and to construct valid new corpora, research teams need some sense of the models that are relevant to the area.

What are appropriate sources?

In the area of emotion, some of the hardest problems involve acquiring basic data. Four main types of source are commonly used. Their potential contributions and limitations need to be understood.

Acted: Many widely used emotion databases consist of acted representations of emotion (which may or may not be generated by actors). The method is extremely convenient, but it is known that systems trained on acted material may not transfer to natural emotion. It has to be established what kind of acted material is useful for what purposes.

Application-driven: A growing range of databases are derived from specific applications (e.g. call centres). These are ideal for some purposes, but access is often restricted for commercial reasons, and it is highly desirable to have more generic material that could underpin work on a wide range of applications.

General naturalistic: Data that is representative of everyday life is an attractive ideal, but very difficult to collect. Making special-purpose recordings of everyday life is a massive task, with the risk that recording changes behaviour. Several teams have used material from broadcasts, radio & TV (talk shows, current affairs). That raises issues of access, signal quality, and genuineness.

Induction: A natural ideal is to induce emotion of appropriate kinds under appropriate circumstances. Satisfying induction is an elusive ideal, but new techniques are gradually emerging.

Which modalities should be considered, in which combinations?

Emotion is reflected in multiple channels - linguistic content, paralinguistic expression, facial expression, eye movement, gesture, gross body movement, manner of action, visceral changes (heart rate, etc), brain states (eeg activity, etc). The obvious ideal is to cover all simultaneously, but that is impractical - and it is not clear how often all the channels are actually active. The community needs to clarify the relative usefulness of the channels, and of strategies for sampling combinations.

What are the realistic constraints on recording quality?

Naturalism tends to be at odds with ease of signal processing. Understanding of the relevant tradeoffs needs to be reached. That includes awareness of different applications (high quality may not be crucial for defining the expressive behaviours a virtual agent should show) and of timescale for solving particular signal processing issues (eg. recovering features from images of heads in arbitrary poses).

How can the emotional content of episodes be described within a corpus?

Several broad approaches exist to transcribing the emotional content of an excerpt - using everyday emotion words; using dimensional descriptions rooted in psychological theory (intensity, evaluation, activation, power); using concepts from appraisal theory (perceived goal-conduciveness of a development, potential for coping, etc). These are being developed in specific ways driven by goals such as elegance, inter-rater reliability, and faithfulness to the subtlety of everyday emotion, relevance to agent decisions, etc. There seems to be a real prospect of achieving an agreed synthesis of the main schemes.

Which emotion-related features should a corpus describe, and how?

Corresponding to each emotion-related channel is one or more sets of signs relevant to conveying emotion. For instance, paralinguistic signs exist at the level of basic contours - F0, intensity, formant-related properties, and so on; at the level of linguistic features of prosody (such as 'tones and breaks' in TOBI); and at more global levels (tune shapes, repetitions, etc). Even for speech, inventories of relevant signs need to be developed, and for channels such as idle body movements, few descriptive systems have been proposed. Few teams have the expertise to annotate many types of sign competently, and so it is important to establish ways of allowing teams that do have the expertise to make their annotations available as part of a database. Mainly for lower level features, automatic transcription methods exist, and their role needs to be clarified. In particular, tests of their reliability are needed, and that depends on data that can serve as a reference.

How should access to corpora be provided?

Practically, it is clearly important to find ways of establishing a sustainable and easily expandable multi-modal database for any sorts of emotion-related data; to develop tools for easily importing and exporting data; to develop analysis tools and application programmers interfaces to work on the stored data and meta-data; and to provide ready access to existing data from previous projects. Approaches to those goals need to be defined.

What level of standardisation is appropriate?

Standardisation is clearly desirable in the long term, but with so many basic issues unresolved, it is not clear where real consensus can be achieved and where it is better to encourage competition among different options.

How can quality be assessed?

It is clear that some existing corpora should not be used for serious research. The problem is to develop quality assurance procedures that can direct potential users toward those which can.

Ethical issues in database development and access

Corpora that show people behaving emotionally are very likely to raise ethical issues - not simply about signed release forms, but about the impact of appearing in a public forum talking (for instance) about topics that distress or excite them. Adequate guidelines need to be developed.

The number and quality of submissions were well above our expectations, out of 26 submitted papers, 10 papers were accepted for oral presentations and 12 for poster presentations. They enable the workshop to cover several dimensions of emotional corpora:

- Music and emotion
- Speech and emotion
- Multimodal behaviour
- Acted, simulated and Real-life emotion
- Portrayed emotion
- Annotation scheme and language of representation

We expect that the output of the workshop will contribute to the study of practical, methodological and technical issues central to developing emotional corpora (such as the methodologies to be used for emotional database creation, the coding schemes to be defined, the technical settings to be used for the collection, the selection of appropriate coders).

Looking forward to an exciting emotional workshop!

Laurence Devillers and Jean-Claude Martin, LIMSI-CNRS, France,
devil@limsi.fr, martin@limsi.fr

Roddy Cowie and Ellen Douglas-Cowie, QUB, UK,
r.cowie@qub.ac.uk, e.douglas-Cowie@qub.ac.uk

Anton Batliner, University of Erlangen, Germany.
batliner@informatik.uni-erlangen.de

A Language-Resources Approach to Emotion: Corpora for the Analysis of Expressive Speech

Nick Campbell

Acoustics & Speech Processing Department,
Spoken Language Communication Research Laboratory,
Advanced Telecommunications Research Institute International,
Keihanna Science City, Kyoto 619-0288, Japan.
nick@atr.jp

ABSTRACT

This paper presents a summary of some expressive speech data collected over a period of several years and suggests that its variation is not best described by the term "emotion". Further, that the term may be misleading when used as a descriptor for the creation of expressive speech corpora. The paper proposes that we might benefit from first considering what other dimensions of speech variation might be of more relevance for developing technologies related to the processing of normal everyday spoken interactions.

INTRODUCTION

Spoken language has been extensively studied through the use of corpora for several decades now, and the differences between the types of information that can be conveyed through written texts and those that are signalled through speech are beginning to be well understood.

The paralinguistic information which is perhaps unique to speech communication, is largely carried through modulations of prosody, tone-of-voice, and speaking style, which enable the speakers to signal their feelings, intentions, and attitudes to the listener, in parallel with the linguistic content of the speech, in order to facilitate mutual understanding and to manage the dynamics of the discourse [1].

The different types of information that are signalled by different speaking styles are also well understood and are beginning to be modelled in speech technology applications. The more formal the speech, the more constrained the types of paralinguistic information that are conveyed.

As an example of one extreme, we might consider a public lecture, where the speaker is (sometimes literally) talking from a script, to a large number of listeners (or even to a recording device with no listeners physically present) and has minimal feedback from, or two-way interaction with, the audience. This type of 'spontaneous' speech is perhaps the most constrained, and most resembles text.

As an example of the other extreme, we might consider the mumblings of young lovers. Their conversation is largely phatic, and the words might carry little of linguistic content but are instead rich in feelings. For them, talk is almost a form of physical contact.

There are many steps along the continuum between these two hypothetical extremes of speaking-style variation. Perhaps they can be distinguished by the ratio of paralinguistic to linguistic content, i.e., the amount of 'personal' information that is included in the speech. The lecture, having almost no personal information and a very high amount of propositional content will result in a very low value of this measure, while the phatic mutterings will score very high.

If we are to collect data that contains sufficient examples of natural spoken interactions along the whole range of this continuum of values, then low-scoring material will prove very easy to collect, but most lovers might object strongly to the suggestion of a recording device intruding into their privacy. Thus, by far the majority of speech corpora that have been used in previous research score very poorly on this scale and as a result the speech that they contain is not very far removed from pure text in its style and content.

A CORPUS OF EXPRESSIVE SPEECH

We need more varied and representative corpora if we are to develop future speech technology that is capable of processing the more human aspects of interactive speech in addition to its propositional content. However, the difficulties of doing this are well known. Since Labov, the presence of an observer (human or device) has been known to have an effect on the speech and speaking style of the recorded subject, and unobtrusive recording is unethical, if not already illegal in most countries. Several approaches have been proposed to overcome this obstacle to future research. This section reports one of them, and discusses some of the conclusions that we reached on the basis of that experience. The JST/CREST Expressive Speech Corpus [2] was collected over a period of five years, by fitting a small number of volunteers with head-mounted high-quality microphones and small mini-disc walkman recorders to be worn while going about their ordinary daily social interactions.

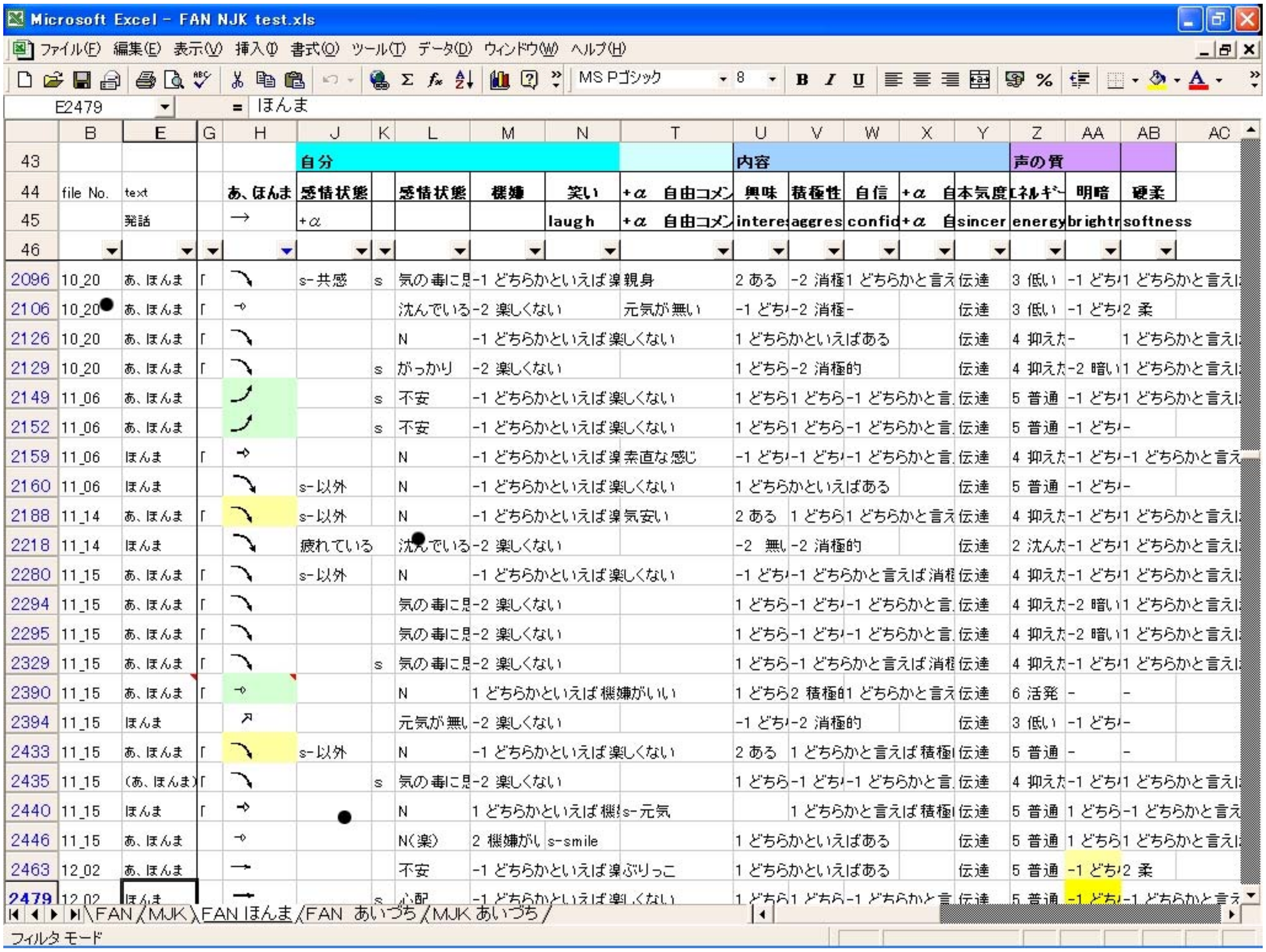


Figure 1: A screenshot of the labelling spreadsheet for the word "honma". The columns include data described in more detail in Table 1. In this form of labelling, tokens are listened to in isolation, free of contextual influence, while in other forms of labelling they are annotated in time-aligned sequence, taking context into account. By clicking on a filename, the labeller can listen to each sample interactively

Further groups of paid volunteers transcribed and annotated the speech data for a variety of characteristics, including speech-act, speaker-state, emotion, relationship to the interlocutor, etc. All the data were transcribed, and about 10% was further annotated. Figure 1 shows a sample of the annotation results, and Table 1 shows some of the categories that were used for annotation. These samples can be listened to at the project web-site, http://feast.atr.jp/non-verbal/. The material is in Japanese, but many of the findings hold for other languages as well. Japanese are people too, and many of the non-verbal speech sounds in this language can be equivalently understood by native-speakers of other languages who have no experience of either the Japanese language or culture. A laugh is a laugh in any language. So is a sigh.

The data in figure 1 represent a few of the approximately 3,500 tokens of the Japanese word /honma/ from one speaker of the corpus. The word functions in much the same way as "really" does in English; both as a qualifying adjective (really hot!) and as an expressive exclamation (really?!). The word is typical of many that are repeated frequently throughout the corpus, and that are used by the speaker more for their discourse effect than for their linguistic or propositional content. No two pronunciations of this word are the same, and each carries subtle affective and interpersonal information that signals many kinds of different states and relationships, as will be described in more detail below.

These words proved most difficult for the labellers to adequately categorise. They function primarily as backchannel utterances, but also serve to display a wide range of attitudinal and affective states. We have reported elsewhere [3] studies that measure the extent to which their function can be similarly perceived by different groups of listeners belonging to different cultural backgrounds and languages. In terms of quantity, more than half of the utterances in the corpus were of this type; short words or simple syllables that occurred alone or were repeated several times in succession, often not appearing at all in a dictionary of the formal language, but forming essential components of a two-way spoken interaction.

ANNOTATING THE CORPUS FOR EMOTION

It is clear that these types of expression carry emotional information. They are very expressive, and revealing of the speaker's type(s) and degree(s) of arousal. We therefore attempted to label emotion in the corpus data.

A version of the Feeltrace software was implemented (square, rather than round!) and each utterance was assigned a value within the valency/arousal space thus defined. The labellers understood the meaning and validity of these two dimensions, and felt easy about working with the mouse-based software for data entry, but most complained about the work after a short time. They claimed that the framework simply wasn't appropriate for describing the different types of variation that they perceived in the speech. They proposed instead the descriptive categories shown in Table 1.

While the speaker was clearly in a given state of emotional arousal during each utterance, the correspondence between what the labellers could determine about the speaker state, from various contextual and expressive clues, and how the speaker's utterance was *performing* in terms of her stance within the discourse, was often very small.

When labelling five-years worth of someone's speech, you become very familiar with that person's mannerisms and even those of their circle of acquaintances. For example, it might be clear from various such clues that the speaker is angry on a given day. Yet the presence or absence of anger in a person may have little or no relationship to the presence or absence of anger in the expression of a given speech utterance. How is this to be labelled in the simple valence/arousal framework?

Specifically, let's examine three such cases:

(i) A schoolteacher walks into the classroom and the children continue to be noisy. The teacher gets angry with the children.

(ii) The same teacher has been wrongly accused of malpractice during the lunchbreak and continues to teach in the afternoon. She explains to the children the details of the lesson.

(iii) The same teacher later in the afternoon as the children persist in being noisy. She gets angry with them again.

In the first case, the speaker expresses anger but does not feel it - she is merely doing her job, and performing an expected role in order to achieve a predictable effect. The children know the rules and soon stop talking. In the second case, the opposite is happening; the person is angry, but her speech is not; as a professional, she continues to speak to the children in the way to which they have become accustomed. In the third case, we have an angry person who is being angry. The effect on the children is immediate. They are afraid.

The three types of speech illustrated above all contain anger, but they differ in whether it is felt or expressed. We could further differentiate by degree of anger, or degree of expression, or both, and with respect to degree of expression, also determine whether "something inside is being let out" or whether the voice is being made to sound as though it is, when in fact inside the feelings may be neutral (whatever that expression might mean).

Affect and Attitude in the Speech

In view of the above, the labellers felt that it was preferable to work with a three-level labelling system, where (i) facts about the speaker could be distinguished from (ii) facts about the speech, and (iii) separate independent evaluations could be made about the information portrayed by the voice. After some experimentation, the system detailed in Table 1 was proposed.

Level 1 describes the state of the speaker, requiring long-term context, and an estimation of the discourse purpose of the utterance (see details below), the speaker's emotion and mood (these labels are free-input, those in the table being examples), her interest in the discourse, and finally a label to denote labeller-confidence. Numerical labels are forced-choice on a scale of high to low (see lower part of table) with no default or zero setting.

Table 1: Three levels of labelling for describing each utterance, including use of six-level forced-choice tendency scales

Level 1	STATE (about the speaker)
purpose	a discourse-act/DA label (see text)
emotion	happy/sad/angry/calm
mood	worried/tense/frustrated/troubled/...
interest	a 6-point scale from +3 to -3, omitting 0
confidence	a 6-point scale from +3 to -3, omitting 0
Level 2	STYLE (about the speech)
type	speaking-style label (open-class)
purpose	a discourse-act label (closed-class)
sincerity	insisting/telling/feeling/recalling/acting/...
manner	polite/rude/casual/blunt/sloppy/childish/sexy/...
mood	happy/sad/confident/diffident/soft/aggressive/...
bias	friendly/warm/jealous/sarcastic/flattering/aloof/...
Level 3	VOICE (about the sound)
energy	a 6-point scale from +3 to -3, omitting 0
tension	a 6-point scale from +3 to -3, omitting 0
brightness	a 6-point scale from +3 to -3, omitting 0
level 0	labeller
confidence	a 6-point scale from +3 to -3, omitting 0

6-point values:	negative	positive
'very noticeable'	-3	3
'noticeable'	-2	2
'only slightly noticeable'	-1	1

Level 2 describes the style of the speech, its type and purpose, and can be estimated from a short-time window (i.e., no context) so that it describes the information available from listening to the isolated speech utterance alone, as distinct from the same utterance situated in a discourse (i.e., we don't care if she is angry or not, but this segment SOUNDS angry). The *sincerity* label describes an important functional aspect of the speech, such as can be distinguished between the verbs 'insisting', 'telling', 'quoting', 'saying', 'feeling', 'recalling', 'acting', 'pretending' etc.

An example from the corpus will illustrate how difficult it can be to assign such apparently simple labels. The speaker, a young woman, says something in Japanese that might translate into: "You're a f***ing pig! I shouted and stormed out of the place!". It was told by the young woman to a sympathetic friend who was laughing with her over the row she and her husband had had the previous evening. On listening to the first few words in isolation, the listener can hear only extreme anger. However, there is no gap in the speech and by the time we reach "stormed out", the speaker is giggling as she speaks, and then finally the utterance ends in real guffaw laughter.

In the example above, we would select 'quoting' (self) rather than 'acting' or 'feeling' for the expletive, and 'feeling' for the laughter at the end, but still have no way to ex-plain the slide of "emotions" (is that the right word?) from start to end of the utterance, which lasted little more than a second. Fortunately, not all utterances are as complex, and most were satisfactorily assigned a single label for each category in the table.

Manner is a bucket category that includes politeness and sexiness (which are not at all mutually contradictory) as well as childishness, sloppiness, etc to describe the perceived attitude(s) of the speaker towards the listener. This is complemented by Mood and Bias, of which the former indicates the affective states of the speaker, and the latter his or her attitudes.

Level 3 describes the physical aspects of speaker's voice quality and speaking style in perceptual terms.

Discourse-Act Labelling

In order to describe the purpose or function of each utterance, a decision was first made about its *directionality*, which may be either 'offering' (to the listener) or 'seeking' (from the listener). Utterances were then broadly categorised into seven classes of *discourse intentions*, including Questions, Opinions, Objections, Advice, Information, Greetings, and Grunts. These category labels were determined by necessity as examples of each appeared in the data; the last category accounted for almost half of the utterances in the corpus.

Under the category of *Questions*, we use the following labels: WH Questions, Y/N Questions, Repetition Requests, and Information Requests.

Under the category of *Opinions* we use the following labels: Opinion, Compliment, Desire, Will, Thanks, and Apology.

Under the category of *Objections* we use the following labels: Objection, Complaint.

Under the category of *Advice* we use the following labels: Advice, Command, Suggestion, Offer, and Inducement.

Under the category of *Information* we use the following labels: Give Information, Reading, Introduce Self, Introduce Topic, and Closing.

Under the category of *Greetings* we use the following labels: Greeting, Talking to Self, Asking Self, Checking Self.

Under the category of *Grunts* we use the following labels: Notice, Laugh, Filler, Disfluency, Mimic, Habit, Response, and Backchannel. Response and backchannel utterances are further subcategorised into the following types: agree, understand, convinced, accept, interested, not convinced, uncertain, negative, repeat, self-convinced, notice, thinking, unexpected, surprise, doubt, impressed, sympathy, compassion, exclamation, listening, and other.

EXPRESSIVE SPEECH AND EMOTION

The experience gained from this labelling process has caused us to now rethink our original assumptions. We started off by trying to overcome Labov's Observer's Paradox, hoping that long-term exposure to a recording device would eventually cause the wearer to familiarise with it to the extent that it no longer becomes a hindrance to normal spoken interaction, even of a highly personal kind. This has proven to be the case, as the variety of speech that we have collected well shows.

However, another paradox has arisen in its place. We originally believed that we would be able to capture truly natural and spontaneous emotional speech data by having a microphone active and in place before and while the emotional 'event' took place. Instead, we find that by far the majority of our speech material is NOT marked for emotion as we then conceived it, but that it varies significantly in dimensions better related to affect and attitude, signalling the mood and interest of the speaker, his or her current relations with the listener, and controlling the variable flow of the discourse.

We started out by believing that 'emotion' was the essential component lacking in our speech corpora for technology development, but we now consider that the 'human dimension' that we were looking for is not best described by the term "emotion" at all. Our data score very highly on the measure of paralinguistic to linguistic content described in the introduction, and are very far from the formal speech of less interactive situations, almost half being nonverbal and affect-related, but they lead us to conclude that the emotional state(s) of the speaker are not always directly expressed, and that social and interpersonal considerations override the supposed link between subjective emotion and displayed affective states. The social aspects of communication take precedence over the blunt expression of feeling, and while the latter can perhaps be determined from an expressive utterance, the multiple levels of information in the former provide a richer source of data to be processed if we are to "better understand the person" through her speech.

CONCLUSION

Since it is of great importance to present experiments with real examples and to have theoretical discussions based on analysis of representative data, it is of fundamental importance to clarify emotional representation, data collection aim and methodology to obtain data. Many corpora of speech are now being designed to maximise the inclusion of emotional samples, so that progress may be made in the understanding of all aspects of human interactions, but because of the difficulty in collecting natural spontaneous materials, actors are being used to simulate the target speaking styles and emotional states. They are undoubtedly very competent and will produce exactly the material that we ask for, but in trying to please us, are they giving us what we really need? In constraining our requests to "emotion" are we not in danger of missing so much more that is perhaps the core of human interpersonal interactions? Our experience with the ESP corpus leads to the conclusion that this might be the case.

REFERENCES

1. Campbell,, N., "Getting to the heart of the matter; Speech as Expression of Affect rather than Just Text or Language", pp 109-118, Language Resources & Evaluation Vol 39, No 1, Springer, 2005.
2. The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.his.atr.jp
3. Campbell, N., & Erickson, D., "What do people hear? A study of the perception of non-verbal affective information in conversational speech", pp. 9-28 in Journal of the Phonetic Society of Japan, V7, N4, 2004.

A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process

Frank Enos and Julia Hirschberg

Department of Computer Science
Columbia University, New York City
frank@cs.columbia.edu

ABSTRACT

This paper offers an approach and a theoretical framework for eliciting emotional speech using actors. The framework is developed by connecting the goal-based model of emotion proposed by Abelson [1], the work of appraisal theorists, and an approach to the actor's technical process widely used in the professional theater and taught in modern conservatories. In doing so, we hope to address some of the difficulties currently encountered in the use of acted speech in emotion research.

Keywords

Emotion, elicitation, acting, appraisal theory.

INTRODUCTION

Numerous studies of emotional speech have employed acted speech with varying degrees of success. We believe that the utility of this approach to elicitation can be augmented by a more thorough understanding of the actor's process and how it relates to emotion. Campbell [3] gives a thorough accounting of the problems presented by the use of acted speech. He details the concern that such speech may be limited in dimension, and, more subtly, that in social interaction, expression of emotion may be unintentional. In contrast, in most studies that use acted speech, the expression of a particular emotion is both intentional and overt. Indeed, many such studies simply instruct the actor to utter a given text "with emotion X" (for a listing of such studies, see [11]; a list of databases can be found in [6]). Notable exceptions include those studies that employ the 'scenario approach' such as [2] and a study currently underway by the Geneva Emotion Research Group [Bänziger, personal communication]; Scherer et al. [10] provide a list.

Even in studies that employ scenarios, however, there is usually overt instruction to the actor with respect to specific emotions. We argue that directing the actor to focus on the goal of encoding a particular emotion works at cross-purposes with the advantages conferred by the trained actor's process. Consider a play in which an actor attempts to convey sorrow by forcing himself to weep. Such a ploy is often unconvincing, and transparent even to a naïve observer. Most importantly, it is generally inconsistent with human behavior. This is an extreme form of the sort of artificial manipulation entailed in many emotional speech elicitation paradigms. It assumes that the actor has been trained – and has achieved the capacity – to manipulate directly the physiological mechanisms involved in the expression of natural emotion. In most cases, this is simply not true.

Preliminary assumptions

For the purposes of this paper, emotion is understood to be defined in the following complementary ways:

1. 'Reaction to significant stimulus events that impinge on organismic equilibrium or that change the organism-environment relationship.' [9]

2. 'Valenced reactions to events, agents or objects, with their particular nature being determined by the way in which the eliciting situation is construed.' [8]

Further, the consideration of acting given here is primarily concerned with approaches to stage acting, in particular, acting in the tradition of the American and British stage. This is because in theater, as in daily life, the perception of emotion on the part of an observer or interlocutor is to a degree unmediated by other forces. In film or television, the director, via the camera and editing, makes crucial decisions with respect to what is observed [McEleney, personal communication], and thus the conveyance of emotion in film depends heavily on a number of factors external to the actor and the observer. Although the stage director controls many elements of the interaction, the contact is ultimately live and direct. Finally, it is not suggested that all actors work in the way described here. The approach described is, however, founded on six years of conservatory training and fifteen years of professional experience on the part of one of the authors, and we have found it to be one effective means of accomplishing the actor's task.

TECHNICAL ASPECTS OF ACTING

We begin with a personal observation: an inexperienced or poorly-trained actor who is having trouble conveying a scene convincingly can often readily report what he or she is feeling. In contrast, a convincing actor will often report what he or she is doing. The latter is consistent with the approaches advocated by many respected teachers of acting, for example Michael Chekhov [4], the faculty of the Brown/Trinity Rep Consortium, and at its' root, Konstantin Stanislavski [14]. In these approaches, (in Chekhov's [4] terminology) the actor develops an 'action' (possibly an intention; sometimes literally an action) for each segment of the script, and, working with the director, develops elements of character (personality) and atmosphere (context) in which to pursue the action. We argue below that this approach has the potential to simulate the appraisal process. This happens in terms of the personality and background of the character, and in terms of the events, agents, and context of the given scene. In turn, this simulation allows the mechanisms of emotional expression — facial, vocal, physical — to engage naturally. This process is complemented by aspects of the actor's training that develop physical and mental responsiveness to emotion inducing-stimuli.

We do not suggest that all approaches to acting are reducible to this approach. We do argue that this approach may lead to more viable and natural elicited emotional speech than has been previously obtained from trained actors.

What (some) actors do

A widely accepted claim in acting is that "Good actors try to do what people do" [McEleney]. What then distinguishes theater from the events of our daily lives? We contend that the appeal of theater is found in the opportunity to observe 'what people do' in extraordinary situations. These situations may be extraordinary because of momentous events such as the shipwreck that precipitates the events of Shakespeare's 'Twelfth Night'; because of an unusual context, such as the enchanted island of 'The Tempest'; or because of the emotionally charged relationships of the characters. A good play often combines these characteristics. This claim can thus be reframed: Good actors, on stage, try to do what people do under extraordinary circumstances. In the best case, such portrayals necessarily generate emotion, since 'human behavior in a high stakes situation' is virtually synonymous with the eliciting circumstance described in Definition (1) of emotion above. Crucially, the intensity of emotion is modulated by what is at stake.

Acting and goals

What exactly is meant here by human behavior? Certainly not 'to have emotion X'. We claim, for example, that it is not a human activity, in the sense of a primary, intentional activity, to be sad. Display rules [7] may dictate that one appear sad when one is actually happy (e.g. a rival fails spectacularly). The operative word, though, is 'appear', since the true emotion experienced is that of happiness; the intentional activity in this case is 'to *appear* X'. [1]

Novice actors often confuse behavior with 'pretend' behavior [Berenson]. The skilled actor does not pretend to pick up a pistol from the stage floor, he simply picks it up. The manner in which he picks it up is dictated by his understanding of what he, the actor, would have to do in order to do what the character does. In this version of what Stanislavski called the 'Magic If' [13], the 'pretend' takes place in the actor's preparation. As he clarifies the imagined circumstances and goals, it becomes increasingly easier to commit to the action. In turn, emotion arises from the interaction of goals, context, and obstacles, which become increasingly believable to the actor via rehearsal. This obviates the need to manipulate the display of emotion.

The trap of focusing on the emotion is often described by actors as 'playing the result' rather than 'playing the action' [Berenson, personal communication]. In focusing on the result, the actor is asking the question 'what would I sound like?' instead of 'what would I *do*?'. The actor tries to manipulate the body, face, and voice in order to indicate based on some preconception. Again, this is not human activity: the actor engages in behavior only an actor would perform.

EMOTION, GOALS AND APPRAISAL

Abelson [1], writing on cognitive consistency theory, proposes what he calls a functionalist model of emotion. His model contains the following components: Goals, Actions (of the subject), Causal Instrumentality (other agents or events), and Outcomes. Emotion arises as a consequence of the goal relationships in a situation, possibly moderated by expectation. For example, disappointment arises when, given a goal, action, and the influence of some agent, a positive outcome is expected, but the result is instead negative. Likewise, frustration arises when, given the individual's goals and actions, a positive outcome might be expected, but is thwarted by the interference of some agent.

It further seems straightforward to relate Scherer's [9] four major classes of appraisal criteria to Abelson's goal-based model, and consequently to the acting approach we offer. These criteria are (briefly): (1) intrinsic characteristics of objects or events; (2) the significance of the event for the individual's needs or goals; (3) the individual's ability to influence or cope with the consequences of the event; (4) the compatibility of the event with social or personal standards. Ableson's model is equally compatible with other extant work in appraisal theory, for example, Clore's [5] Immediacy Principle: 'Affective feelings tend to be experienced as reactions to current mental content.' And one reading of this principle suggests that the more thoroughly the actor commits to her beliefs with respect to the character, the closer she comes to producing actual emotion.

Although not fully developed here, this connection of appraisals with goals may help to clarify why a goal-based approach could be effective in producing more

[1] Granted, it is not an uncommon goal 'to be happy', but this entails a more existential scope than does the corresponding emotional state.

realistic emotion than is usually achieved by actors in the laboratory. Since some appraisal criteria – such as beliefs regarding the characteristics of objects or the significance of events with respect to goals – may be influenced by imagination, the actor's work may simulate appraisal criteria in a manner that allows emotion to engage and arise naturally.

TWO SUGGESTED APPROACHES

The task of designing an emotion elicitation experiment is, in a sense, to work backwards. One can deconstruct the desired emotion into component parts – goal or intention, context, and text – which the actor then 'reassembles' into emotion, spoken and otherwise. We offer two approaches to eliciting emotion in this fashion. The first revisits methods used by other researchers, but with a focus on the component parts described here. The second makes direct use of the process and materials most familiar to actors.

The Scenario Approach

A number of studies take this approach (e.g. [2, 10]), and it is highly compatible with the methods proposed here. This is even more true for those paradigms that develop scenarios in concert with the actors, possibly via improvisation, as in a recent data collection effort by the Geneva Emotion Research Group [Bänziger, personal communication]. We add to approaches already in use the idea of reverse-engineering the desired emotion into the components most useful to actors. Such an eliciting scenario is composed of:

1. A description of the character.
2. A clear description of the situation, including obstacles to the goal, and what is at stake.
3. A (single) goal for the actor.
4. A text that is meant to accomplish the goal, regardless of whether the specific lexical content embodies that goal.
5. Description(s) of additional character(s) involved, if any.

Below is an example scenario for the emotion anger:

1. The speaker's name is June. She is the mother of three children living in a small American town in the state of Pennsylvania. She works during the day as a waitress in an inexpensive restaurant. Most nights, she cleans the offices of a man who is a customer of the restaurant. Most nights, the offices do not need to be cleaned. The owner has given her the job because he feels sorry for her. She knows this. One year prior to the scenario, her husband Bill went to the corner store to buy a lottery ticket and never returned. Her family lives in Arkansas; she left Arkansas with Bill seven years earlier to move to New York, where he planned to work as a professional poker player. Pennsylvania was the closest they came to New York. Bill worked for a time in his second cousin's an auto repair shop.
2. June has just returned home at midnight on a week night from cleaning offices to find Bill sitting in the living room playing with the children, cigarette in his mouth and can of beer in his hand. The children worship him like a conquering hero returning from battle.
3. To make her husband leave her home.
4. 'Don't even speak to me!'
5. Bill, as described above, smug and unapologetic.

The level of detail in this example may be *less* that that desired in an actual elicitation scenario. It is important to understand that the average professional stage production rehearses for six weeks or more. In such a production, good actors, often assisted by the playwright, director, and other actors, develop a very specific picture of the life and world of their characters. This may include details not superficially relevant to the action of the play: what foods the character dislikes; what magazines she reads; what she wears to sleep. The more the actor 'knows' about the character, the more readily she can commit to her behavior.

The five elements can be viewed as an equation, the output of which is human behavior, including emotion [Berenson]. The work of the researcher, assisted by a skilled director, is to tune the parameters of this equation in order to elicit the desired result. This should happen *without mention to the actor of the desired emotion*. The charge to the actor is to accomplish some action by uttering the text in such a way that it effects the goal; the actor's focus must be on action.

If, for example, the result is not sufficiently angry for the purposes of the elicitation, rather than focusing the actor on this result, the director might raise the stakes: 'Imagine now that Bill has been gone for three years.' Or that Bill laughs and begins to speak. Perhaps the goal is cold anger: 'Imagine that Bill has been gone for 10 years.' Campbell [3] raises the valid concern that acted emotion is too overt. Perhaps June wishes to make Bill leave without upsetting the children – how does she effect this? The subtext of anger will remain, but the focus is on what June is *doing*; the actor allows the emotion to take care of itself. A primary tool for doing so is specificity [McEleney]. The rehearsal process that fills out the life, the world, and the personality of the character is a process of making clear, specific choices. A scenario can be filled out by embedding the target utterance in a scene, or actors can improvise so that they can more thoroughly commit to the events.

We suggest two possible methods for a priori validation. First, scenarios can be rated for emotional content by a group of judges who can be asked to determine the emotion they would likely experience as the character(s).

In this way, one can have some assurance beyond intuition that the given scenario is likely to elicit the desired emotion. Second, a more data-driven approach to scenario development might be taken (as in e.g. [10]), in which scenarios are constructed as above but based on an existing database of actual subject experience. One such a database is the International Survey On Emotion Antecedents And Reactions (ISEAR) [12], which details the emotional experiences of thousands of cross-cultural subjects.

The script approach

A second proposed method of elicitation seeks to make direct use of the actor's training and skills. In this approach, existing, well-crafted scripts are used, in their entirety or in the form of individual scenes, as the scenarios of elicitation (scenes may be validated a priori as above). If actual language is to be used, it is preferable to select a script that uses contemporary language; certainly language that is likely to be familiar to any subjects who may label the speech. If pseudo-linguistic content (see below) is to be used, the script may be selected from any that are approachable by the actors employed, since it will serve only to develop the actions and events that elicit the emotion. Development of scenarios proceeds as previously described: the five points mentioned above are developed with respect to each scene of interest, and the text is chosen from the play itself. This approach has the advantage that a well-crafted play is literally designed for the actor's methods. The characters, events, and goals will present themselves, with the help of a director, through the process of rehearsal. Take for example a scene from *Romeo and Juliet* of Shakespeare (chosen for its familiarity; see note below on language). The goal here is to elicit from Romeo the emotion 'despair' as he conveys to Friar Lawrence his dismay at having been banished for killing the character Tybalt.

> *Hadst thou no poison mixed, no sharp-ground knife,*
> *No sudden mean of death, though never so mean,*
> *But 'banished' to kill me? – 'banished'?*
> *O friar, the damned use that word in Hell;* [13]

Here the script provides a goal (to be with Juliet), a context (banishment, never to see her again), and a well-developed atmosphere and set of characters. Tokens can be selected from the text, and the actors coached as desired.

Both these approaches are compatible with elicitation scenarios that employ pseudo-linguistic content, but it is crucial to avoid the actor's imposition of preconceptions with respect to the 'sounds' of emotions. Again, the solution is to approach the speech from the standpoint of intention or goals. With the archaic language of Shakespeare, the actor develops an intention and a set of internal and external circumstances so specific that she could only express them in the language of the script [McEleney]. Analogously, the actor must imbue pseudo-text with an intention so specific that the lack of semantic meaning is inconsequential.

CONCLUSIONS

This paper has described an approach to eliciting acted emotional speech that makes use of the trained actor's process. This approach avoids conveying specific emotional labels to the actor in favor of providing goals and contexts in which to play out certain actions. It is suggested that in this manner the actor can engage the mechanisms of appraisal, and thereby produce emotion in a more natural way. The authors have not addressed the process of post-hoc validation of the elicited emotion. It should also be mentioned that this approach assumes the involvement of skilled professional actors; the outcome of using this approach with college or amateur actors is likely to be disappointing. Finally, we think two areas merit further consideration. First, this approach may also be useful in studies of facial expression or body posture (e.g. [15]). Finally, we find the interaction between goals and appraisal interesting, in the context of acting and otherwise.

ACKNOWLEDGMENTS

We thank Tanja Bänziger for information about her current work; and Marc Schröder for comments on some of the ideas presented here. We are grateful to Stephen Berenson and Brian McEleney of the Brown University / Trinity Repertory Theater Consortium for helpful conversations about this work. This research was funded in part by NSF IIS-0325399 and the Department of Homeland Security.

REFERENCES

1. Abelson, R.P. What Ever Became of Consistency Theory? Pers. and Social Psych. Bull. 9(1), 1983, 37-54
2. Banse, R. & Scherer, K.R. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3), (1996), 614-636.
3. Campbell, N. Databases of emotional speech. Proc. ISCA Workshop on Speech and Emotion, (2000) 34-38.
4. Chekhov, M. To the Actor, Harper & Row, NYC (1953).
5. Clore, G. L., Gasper, K., and Garvin, E. Affect as Information, in J. P. Forgas (ed.), Handbook of Affect and Social Cognition, Lawrence Erlbaum Associates, Inc., Mahwah, NJ (2001).
6. Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. Emotional speech: towards a new generation of databases. Speech Communication, 40, (2003) 33–60.
7. Ekman, P. Universals and cultural differences in facial expressions of emotion. In J. Cole (ed.), Nebrasaka Symposium on Motivation, 1971. University of Nebraska Press (1972), 207- 283.
8. Ortony, A., Clore, G. L., and Collins, A. The Cognitive Structure of Emotions, Cambridge University Press, Cambridge, UK, (1988).

9. Scherer, K. R. Appraisal theory. In T. Dalgleish, & M. Power (Eds.). Handbook of Cognition and Emotion (pp. 637-663). Chichester, (1999).

10. Scherer, K. R., Banse, R., & Wallbott, H.G. Emotion inferences from vocal expression correlate across languages and cultures. Journal of Cross-Cultural Psychology, 32(1), (2001), 76-92

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

12. Scherer, K. R., & Wallbott, H.G., Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology, 66, 310-328, (1994).

13. Shakespeare, W. Romeo and Juliet, The Riverside Shakespeare, Houghton Mifflin, New York, (1974).

14. Stanislavski, K. An Actor Prepares, Hapgood, E.R., trans., Theater Art Books, New York, 1971.

15. Wallbott, H. G. Bodily expression of emotion. European Journal of Social Psychology, 28, (1998), 879-896.

16. Bänziger, T. personal communication, March 30, 2006.

17. McEleney, B. personal communication, March 21, 2006.

18. Berenson, S. personal communication, March 26, 2006.

A Coding Scheme for Designing: Computational Model of Emotion Elicitation

Magalie Ochs[1, 2], Catherine Pelachaud[2], David Sadek[1]

[1] France Telecom, R&D Division, Technology Center, France,
{magalie.ochs, david.sadek}@francetelecom.com
[2] LINC Lab., IUT of Montreuil, Université Paris VIII, France,
c.pelachaud@iut.univ-paris8.fr

ABSTRACT

In this article, we present a coding scheme that enables one to annotate dialogs that lead to emotions. It is constructed from a study of sequences of events that triggered user's emotions in real vocal human-machine dialog corpora. Its analysis is based on appraisal theory of emotions. The coding scheme is specifically designed to easily integrate the results of the annotation dialog analysis in a rational agent.

Keywords

Emotion elicitation, coding scheme, rational dialog agent.

INTRODUCTION

A growing interest in designing animated characters expressing emotions has been observed in recent years. This is motivated by an attempt to enhance human-machine interaction. To be able to express emotions, an agent has to know the circumstances under which emotions are triggered. According to cognitive appraisal theories (Scherer, 2000), emotions are elicited by the subjective evaluation of an occurred or anticipated event. Most of existing computational models are based on these theories and more particularly on the well-known OCC model (Ortony *et al.*, 1988). Researchers (El-Nasr *et al.*, 1998; Rosis *et al.*, 2003) have particularly focused on the emotions related to a person's goals (called the goal-based emotions) (Ortony *et al.*, 1988). Generally, animated characters express a positive emotion when an occurred event enables one to achieve one's goal or to increase its achievement probability; they express a negative emotion in the opposite case (Reilly, 1996; El-Nasr *et al.*, 1998; Rosis *et al.*, 2003). Let us imagine an animated character a user can consult to find information on a specific domain. The agent's goal is to provide the user with the requested information. To be believable, the agent should not express a positive emotion after each achieved goal. Some researchers (Breazeal, 2003; DuyBui, 2004) have proposed to use a threshold. The emotion is then expressed only if its intensity is superior to the threshold. However, an agent that expresses emotions after an arbitrary number of good or bad responses is somehow unbelievable. We aim to identify under which circumstances a virtual agent should express emotions to be believable. To achieve this goal, we propose to analyse the sequences of events that lead to user's emotions in real vocal human-machine dialog corpora. In this paper, we present a coding scheme to annotate human-machine dialogs that lead to emotion. The coding scheme is specifically designed to easily integrate the results of the annotation dialog analysis in a rational agent. Before describing the dialog corpora in more details, we present the model of rational dialog agent we use. Finally, we introduce the coding scheme.

Model of Rational Agent

To create virtual dialog agent, we use a model of rational agent based on a formal theory of interaction (called *Rational Interaction Theory*) (Sadek, 1991). This model uses a BDI-like approach. The implementation of this theory has given rise to a rational dialog agent technology (named *Artimis*) that provides a generic framework to instantiate intelligent agents able to engage in a rich interaction with both human interlocutors and artificial agents (Sadek *et al.*, 1997). The mental state of a rational agent is composed of two mental attitudes: belief and intention, formalized with modal operators. Based on this mental state, a rational agent acts to realize its intentions. The corpora we used are composed of dialogs between users and such rational agents. In the following section, we describe in more details their content.

Corpora Description

The corpora we use are derived from two vocal applications developed at France Telecom R&D: *PlanResto* and *Dialogue Bourse*. *PlanResto* is a restaurant guide for Paris. Users interact using natural language with an embodied conversational agent to find a restaurant (for more details see (Ancé *et al.*, 2004)). *Dialogue Bourse* is a stock exchange application that allows a user to consult her accounts, to know rates and to perform transactions.

Corpora are composed of a large amount of non-acted user-rational agent dialogs (that correspond approximately to 1000 dialog turns). They have been annotated with two labels of emotions (positive and negative). The annotations have been done based on vocal and semantic cues of user's emotions. Finally, in the dialogs transcribed in text, these tags represent the moment where user expresses positive or negative emotion.

A Coding Scheme for Annotating Conditions of Emotion Elicitation

The objective of the annotation is to analyse the sequence of events that leads to the expression of a user's emotion. According to the OCC model (Ortony *et al.*, 1988), three groups of emotions are distinguished: the goal-based-emotions (emotions triggered by the occurrence of an event that affects agent's goal, like joy or disappointment), the standard-based emotions (emotions triggered by the occurrence of an event that affects agent's standards, as for example proud or shame), and preference-based emotions (emotions triggered by the occurrence of an event that affects agent's preferences, as for example like or dislike). In our work, we currently focus on *goal-based emotions*. Indeed, they are the most likely to be triggered in the context of an agent a user can consult to find out information.

In a dialog, an event corresponds to a *communicative act*. Researchers have observed that language is not only used to describe something or to give some statement but also to do something with intention, *i.e.* to act (Austin, 1962; Searle, 1969). Then, a communicative act (or speech act) is defined as the basic unit of language used to express an intention. Based on the Speech Acts Theory (Austin, 1962; Searle, 1969), we assure that goals of the user are to achieve the perlocutory effects of the communicative act performed. The perlocutory effects are the intentions that the user wants to see realized in consequence of the communicative act performed. For instance, the perlocutory effect of the act *to inform an agent j of a proposition p* is that the agent *j* knows the proposition *p*.

In order to identify the sequence of events that potentially elicited emotions, the coding scheme should enable to describe the features of the constituent events. Researchers in Cognitive Sciences (Scherer, 2000; Lazarus, 2001; Roseman, 2001) have highlighted the different characteristics of events that are particularly determinant to trigger an emotion. They correspond to *appraisal criteria* (called also *appraisal variables*) that individuals use to evaluate emotion eliciting-events. Concerning the goal-based emotions elicitation, an event is relevant (i.e. that can potentially trigger emotions) only if it affects a person's goal. (Lazarus, 2001). The relationship between the event and the goal (such as its threatening or its achievement) determines the type of emotions. Moreover, the causes of an event (that may be for example another agent) constitute an important appraisal criterion. Finally, the individual potential to cope with the consequences of an occurred event is crucial in emotion elicitation (Scherer, 1988).

In order to identify the different kinds of relationship of an event with user's goals, types of event causes, and coping behaviours, which can occur in human-machine interaction, we have analysed 10 dialogs (that correspond approximately to 1000 dialog turns) extracted from the corpora described in the previous section.

From the dialogs analysed, we have defined the following tags:

- **Relationship that an event can have with a user's goal;** an event is annotated by this tag if:
    - ***Goal achievement tag***: it has enabled the user to achieve a goal that he thought to see realized by this event.
    - ***Goal failure tag***: it has not allowed the user to achieve his goal that he thought to see realized by this event.
    - ***Partial goal failure tag***: it has allowed the user to achieve only partially his goals that he thought to see realized by this event.
    - ***A posteriori goal failure tag***: after this event the user realizes that his goal has not been achieved though he thought it had.
- **Types of causes of an event that lead to a goal failure;** an event is annotated by this tag if:
    - ***Unfeasibility tag***: the agent does not have the capacity to achieve this goal.
    - ***Belief conflict tag***: the agent believes that the user has another goal than his own.
    - ***Goal conflict tag***: the agent had a goal that the user thought to be already achieved.
    - ***Unshared goal tag***: the agent had a goal that the user has not.
- **Coping behaviours that can be adopted by the user after the failure goal;** an event is annotated by the tag if it has allowed the user:
    - ***Goal giving up tag***: to give up his goal.
    - ***Goal adoption tag***: to adopt a new goal.

These tags represent the coding scheme for annotating the corpora. For the annotation, we suppose that the user knows the perlocutory effects of the communicative act and that he has the intention, by performing it, to realize these effects. An example of a dialog annotated (translate in English) follows:

Welcome to PlanResto. PlanResto enables you to look for a restaurant according to its speciality, its localization or your budget.
What are you looking for?
User: the Hippopotamus restaurant euh in 14th urban district
Agent: A restaurant in 14th urban district. I have found more than a hundred. You can precise a speciality, for example specific type of food or budget. You can speak now.

<*Goal* (of having information on the Hippopotamus Restaurant) *failure*>
User : euh
Agent: A Greek restaurant.
<*Belief conflict* (the agent believes that the user has the intention to have information on Greek restaurant>
User: no, grilled food.
Agent: I now have 8 restaurants. You can ask to consult them. What sort of budget are you looking for?
User: euh + fifteen euros
<*Goal adoption* (on the restaurant budget)>
Agent: I have not understood you.
<*Goal failure*>
User: less than twenty euros
<*Goal adoption*>
Agent: A Greek restaurant in the 14th urban district for less than twenty euros.
<*A posteriori goal failure* (the user's goal to inform the agent on his intention to find a grilled restaurant has failed) >
User: cancellation
<*Goal giving up*>
Agent: I have not understood you.
<*Goal failure*>
User: no, not euh a Greek restaurant, a grilled restaurant
Agen : I'm not sure to have understood you.
<*Goal failure*>
User: A restaurant that serves grilled food
Agent: A restaurant in the 14th urban district for less than 20 euros serving grilled food.
<*Goal* (of the user to inform the agent on his intention) *achievement*>
User: yes
Agent: I now have twenty. The first one is called Café des Arts.
User : next
Agent: The next restaurant. The second restaurant is named l'Entrepôt Restaurant. An advice. To switch from one restaurant to the other you can say previous or next.
<*Unshared goal* (the agent had the goal to inform the user on the use of the command next, but the user already knew about it)>

EXPRESSION OF NEGATIVE EMOTION

According to the coding scheme that we have introduced, the dialogs have been annotated by one coder. To validate intra-coder annotation, we project to re-annotate them in one month by the same coder.

The tags that composed the coding scheme are described in terms of mental attitudes (with the modal operators of belief and intention). This semantically grounded formal representation will enable one to easily integrate the results of the analysis of the annotated corpora in a rational agent.

CONCLUSION

In this paper, we have presented a coding scheme that enable one to annotate the sequences of events that lead to emotions in human-machine dialogs. Our final objective is to use the result of the annotation to create a computational model of emotion elicitation for rational agent based on real human-machine interaction and on appraisal theory of emotions. The agent will express emotion similarly to the user. Evaluation of the model will be done to validate the expressions of emotions of the agent.

REFERENCES

1. Ancé, C., Bretier, P., Breton, G., Damnati, G., Moudenc, T., Pape, J.-P. L., Pele, D., Panaget, F. and Sadek, D. "Find a restaurant with the 3D embodied conversational agent Nestor". In: *Proceedings of 5th workshop on Discourse and Dialogue (SigDIAL)*, Boston, USA. 2004.
2. Austin, J. L. *How to do things with words*, Oxford University Press, 1962.
3. Breazeal, C. "Emotion and sociable humanoid robots". *International Journal of Human-Computer Studies*, 59, 119-155.2003.
4. DuyBui, T. Creating Emotions And Facial Expressions For Embodied Agents. Thèse, Department of Computer Science, Univeristy of Twente, 2004.
5. El-Nasr, M. S., Ioerger, T. R. and Yen, J. "Learning and Emotional Intelligence in Agents". In: *Proceedings of AAAI (American Association for Artificial Intelligence) Fall Symposium on Emotional Intelligence*, Florida, USA. 1998.
6. Lazarus, R. S. "Relational Meaning and Discrete Emotions". *Appraisal Processes in Emotion: Theory, Methods, Research*. K. Scherer, A. Schorr and T. Johnstone, Oxford University Press, 37-69. 2001.
7. Ortony, A., Clore, G. L. and A., C. *The cognitive structure of emotions*, Cambridge University Press, 1988.

8.Reilly, S. Believable Social and Emotional Agents. Thèse, informatique, Université de Carnegie Mellon, 1996.

9.Roseman, I. J. "A model of appraisal in the emotion system". *Appraisal Processes in Emotion: Theory, Methods, Research*. A. S. Klaus Scherer, Tom Johnstone, Oxford University Press, 68-91. 2001.

10.Rosis, F. d., Pelachaud, C., Poggi, I., Carofiglio, V. and Carolis, B. D. "From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent". *International Journal of Human-Computer Studies*, 59 (1-2), 81-118.2003.

11.Sadek, D. Attitudes mentales et interaction rationnelle: vers une théorie formelle de la communication. Thèse, informatique, Université de Rennes I, 1991.

12.Sadek, D., Bretier, P. and Panaget, F. "ARTIMIS: Natural Dialogue Meets Rational Agency". In: *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI'97)*, Nagoya, Japon, 1030-1035. 1997.

13.Scherer, K. "Criteria for emotion-antecedent appraisal: A review". *Cognitive perspectives on emotion and motivation*. V. Hamilton, G. H. Bower and N. H. Frijda, Dordrecht: Kluwer, p 89-126. 1988.

14.Scherer, K. "Emotion". *Introduction to Social Psychology: A European perspective*. M. Hewstone and W. Stroebe, Oxford, 151-191. 2000.

15.Searle, J. R. *Speech Acts*, Cambridge University Press, 1969.

GEMEP – GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions

Tanja Bänziger
University of Geneva
40, Boulevard du Pont-d'Arve,
1205 Geneva, Switzerland
Tanja.Banziger@pse.unige.ch

Hannes Pirker
Austrian Research Inst.
for Artificial Intelligence
Freyung 6, A-1010 Vienna, Austria
Hannes.Pirker@ofai.at

Klaus R. Scherer
University of Geneva
40, Boulevard du Pont-d'Arve,
1205 Geneva, Switzerland
Klaus.Scherer@pse.unige.ch

ABSTRACT

This paper introduces the GEMEP (Geneva Multimodal Emotion Portrayals) corpus, a new repository of portrayed emotional expressions. Corpora of acted portrayals tend to include portrayals of very intense emotions, which are considered to occur infrequently in daily interactions between humans or between humans and machines. Acted portrayals have therefore been challenged as unsuited for applied research purposes. Taking a different stance, we argue that: (a) portrayals produced by appropriately instructed actors are analogue to expressions that do occur in selected real-life contexts; (b) acted portrayals – as opposed to induced or real-life sampled emotional expressions – display the most expressive variability and therefore constitute excellent material for the systematic study of nonverbal communication of emotions. We describe the guidelines used to record the corpus and some of our short-term research plans with the corpus.

Keywords

Emotions, actors, acted portrayals, multimodality, facial expressions, vocal expressions, gestures

INTRODUCTION

Most studies in the field of nonverbal communication of emotion have been performed either with acted or with posed emotional expressions (called *emotion portrayals* in this paper). Classical studies of facial expressions used photographs of people displaying selected "facial action units" (activations of single muscles or groups of muscles) and scored them for emotional expression [5]. For vocal expressions, Juslin and Laukka [3] reported that 87% of the studies they reviewed used acted portrayals. But, despite their over-representation in this research field, emotion portrayals were also always criticized for not being representative of "natural" emotional expressions [2]. Why then did several generations of researchers persist in studying them?

The most usual answer to this question lies in a number of limitations of the two alternatives to portrayals: expressions recorded under controlled conditions created to induce emotions; and expressions recorded in sampled real-life situations [6]. For practical and ethical reasons, emotions induced in experimental contexts are seldom very intense and, correspondingly, expressions in these contexts are scarce and, when they occur, faint. Expressions sampled in real-life occur in specific contexts which often cannot be reproduced, their verbal content and the overall quality of the recordings cannot be controlled, one person is usually recorded in only one or very few different emotional states, and it is not always clear on which grounds and how these states should be labelled. Hence, emotion portrayals offer a number of advantages which are difficult to obtain with either induced expressions or real-life sampled expressions: the possibility to record multiple emotional expressions produced by the same senders (actors); the possibility to obtain strong emotional expressions; and controlled recording conditions, including: uniform lexical content in speech, good acoustic and visual quality, and a clear definition of the expressive intention of the sender. Keeping identity of the sender or verbal content constant across emotions is essential to allow comparison between emotional expressions. Nevertheless, it can still be objected that this sort of control is not desirable if it is detrimental to the object of study, in this case the "natural" emotional expression and/or its context of occurrence. In the following some arguments that have been raised against the usage of emotion portrayals are discussed in more detail.

OBJECTIONS TO EMOTION PORTRAYALS

There are several common objections to the use of emotion portrayals. The classification we propose here is arbitrary but we believe that it accurately describes the main concerns that have been voiced, or may be voiced, regarding the use of portrayals in research.

Portrayals reflect stereotypes, not genuine emotions

Emotional expressions produced by actors reflect cultural stereotypes – maybe even stereotypes pertaining to specific acting schools or acting traditions. A more or less explicit assumption is that portrayals are exaggerated (over-acted) as compared to spontaneous expressions. They convey the expressive intention of the actor but not in a "realistic" way. The receiver will recognize the intention but will not necessarily believe that the sender is genuinely emotional. In cases where actors attempt to display expressive signs without invoking emotional feelings, portrayals might indeed be produced in the absence of genuine emotions. An assumption in this view is that actors who do this might be unable to mimic (simulate or fake) the more subtle signs habitually related to the emotion. In this perspective, without being necessarily exaggerated, the portrayals would lack essential features of genuine expressions.

Portrayals represent infrequent emotions

Actors have been traditionally requested to portray so-called "basic emotions", identified by labels such as: 'anger', 'fear', 'sadness', 'happiness', or 'disgust'. Taking a closer look at the literature it becomes apparent that most studies in which acted material was used were set to study rather "extreme" emotional states, which might be better labelled: 'rage', 'panic', 'depression', 'elation', or 'repulsion'. This range of very intense emotions is considered by some authors to be unrepresentative of the range of emotions that are likely to occur on a daily basis in ordinary interactions [1]. This further led to the conclusion that portrayals are not representative of the range of emotional states that are of interest in specific applied contexts and therefore unsuited for research in corresponding fields (such as human-machine interactions).

Portrayals are decontextualized

Portrayals are produced under conditions designed to remove contextual information, in order to record variations that can be related exclusively to the emotions portrayed. Senders are recorded under identical conditions while they portray a range of emotional states. They are typically seated in front of a uniform background, facing a (video) camera. In instances were vocal or dynamic facial expressions are recorded, the actors are usually requested to first appear inexpressive, then to produce an emotional expression, and finally to get back to their inexpressive "baseline". If vocal expressions are of interest to the researchers, the senders will usually be requested to portray emotions while pronouncing the same sentence in all conditions. While most researchers have favoured this type of controlled recording design, it is certainly true that receivers presented with a relatively large number of such portrayals will not be inclined to perceive them as "natural".

ARGUMENTS FOR EMOTION PORTRAYALS

Acted portrayals allow to record strong emotional expressions and allow comparisons across emotions without systematic confounds, while the same is more difficult to achieve respectively with induced expressions and "naturally" occurring expressions. In the following paragraphs we address the objections raised in the previous section.

Portrayals need not be stereotypical nor faked

Actors can and should be encouraged to produce believable expressions by using acting techniques that are thought to stir genuine emotions through action [4]. When portraying emotions, actors should not exaggerate or fake expressions but rather attempt to reactivate emotional experiences while and through acting.
In everyday life, emotional expressions are directed to receivers with different degrees of intentionality. Some expressions might be truly "spontaneous", not directed or intentionally regulated to have an impact on a receiver; whereas acted portrayals are by definition produced intentionally and directed to a receiver. Processes underlying intentional regulation of emotional expressions and their actual effects on expressions are not well known. It would be worthwhile to systematically investigate the similarities and differences of emotional expressions produced more or less intentionally (in everyday life and/or in the laboratory). This could involve comparing acted portrayals with less "controlled" expressions, recorded under conditions that would not promote emphasis or suppression of expressions for the benefit of a receiver.

Portrayals need not to represent basic emotions only

While it is necessary to use descriptors to distinguish and conceptually organize emotional states, descriptors corresponding to "basic emotions" are probably too broad to offer a useful classification for the study of emotional expressions. For vocal expressions in particular, it has been suggested that the failure to reproduce acoustic profiles reported for a given category – for example 'fear' – across studies might derive form variations on the level of the definition of that category [6]. In different studies emotions labeled 'fear' might correspond to very different states ranging from 'apprehension' to 'panic', which obviously would result in different expressive patterns.

Acted portrayals should clearly not be restricted to "basic emotions". Actors can be requested to portray states that are of interest in specific research contexts (for example "frustration" in response to a malfunctioning device or computer application). Researchers need to have clear aims and correspondingly clear operational definitions of the emotional states of interest in their field. In interactions between humans and computers, very intense reactions might be triggered (for instance when playing a challenging game), but more subtle reactions might be of importance as well (e.g. "frustration"). Portrayals produced by appropriately instructed actors could reflect a variety of "realistic" states in different contexts, the selection and definition of the states portrayed pertains to the researcher.

Add more "context" to emotion portrayals

In most cases, emotional reactions are tightly linked to a specific context of occurrence (a situation/event that triggers the reaction). Portrayals would probably appear more "natural" if they were not produced in the complete absence of eliciting events. In everyday life, it is quite rare that a calm, inexpressive person suddenly becomes very emotional, without any apparent reason, and within a few seconds recomposes herself, appearing perfectly calm and inexpressive again. Furthermore, it is undoubtedly beneficial to formulate precise operational definitions of the emotions to be portrayed. This should include at least a rough description of the situation/event in which the reaction takes place. Hence the minimal "context" defined for a given portrayal could be composed of a scenario describing the situation in which the emotion is elicited and – as it is probably more easy for an actor to direct the portrayal at a designated receiver – a brief interpersonal interaction taking place in this situation.

For the construction of any corpus featuring emotional expressions it is essential to clearly define the range of emotions included. Many objections raised against emotion portrayals would probably lapse if the portrayed states would be better selected and clearly defined. For instance, the notion that acted portrayals are lacking "naturalness" (are stereotyped or exaggerated) might be largely derived from the insistence on recording portrayals reflecting extreme emotions.

PROCEDURE USED FOR RECORDING THE CORPUS

This section outlines some of the guidelines applied to record the GEMEP, with an emphasis on the aspects already introduced in the previous sections.

Selection of portrayed emotions

The affective states portrayed were partly selected to match the states frequently studied in the literature dealing with facial and/or vocal expressions of affect. Some less frequently examined states were also included in order to address specific research questions. A relatively large number of positive states – such as 'pride', 'amusement', 'elation', 'interest', 'pleasure', or 'relief' – was for example included in order to challenge the traditional view according to which only one rather undifferentiated positive state ('happiness') can be reliably communicated via facial cues. In a similar attempt, some states corresponding to the same *family* of emotional reactions were included with various arousal levels (e.g. 'irritated' and 'enraged' *anger*; 'anxious' and 'panic' *fear*). This fulfils at least two aims: (1) Reviews of studies describing acoustic profiles of emotional expressions have recurrently reported differences in acoustic features of vocal expressions mostly related to arousal level, the crossing of arousal level and emotion category should allow to partly disentangle the influence of arousal level and emotion family on vocal expressions. (2) The inclusion of more than one type of *anger* (or *fear*) should result in increased variability of the expressions portrayed and should allow to include a range of variations that are more likely to occur in daily interactions, under the assumption, for example, that 'irritation' occurs more frequently than 'rage' and 'anxiety' more frequently than 'panic fear'.

A further attempt to increase the variability of the expressions was undertaken by requesting the actors to produce some of the emotions with less intensity and with more intensity than the intensity that they thought corresponded to the most 'usual' intensity for a given emotion. An underlying assumption (which remains to be tested) is that the portrayals produced with less intensity might be closer to expressions that could occur in daily interactions, while the portrayals with more intensity might be more exaggerated (or more "stereotypical"). To this "regulation" of the intensity of portrayed states, we added a further request to partially mask some of the expressions (i.e. to portray a relatively unsuccessful deception attempt for some of the affective states).

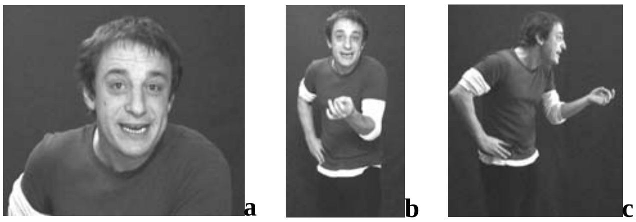
Definition of emotions and of their "context"

Short definitions of the emotional states and "scenarios"[1] were provided to the actors several weeks before the recordings took place. Three "scenarios" were created in order to instantiate each affective state. A "scenario" includes the essential features of a situation, which is assumed to elicit a given emotional reaction. Whenever possible, the scenarios included explicit references to one or more interaction partner(s). The actors were requested to improvise interactions with the director, in which they expressed a given affective state while pronouncing two pseudo-linguistic sentences (1. "ne kal ibam sud molen!"; 2. "kun se mina lod belam?"). The actors were further requested to express each affective state while uttering a sustained vowel, which allowed recording brief emotional expressions in the absence of articulatory movements. For each affective state, the director and the actors were trying different "scenarios" and – after a period of "rehearsal" – recorded one or more interactions until they were satisfied with their performance.

Technical aspects and description of the corpus

Ten professional French-speaking actors portrayed 15 affective states under the direction of (and in interaction with) a professional stage director. Three digital cameras were used for simultaneously recording: (a) facial expressions and head orientations of the actors, (b) body postures and gestures from the perspective of an interlocutor, (c) body postures and gestures from the perspective of an observer standing to the right of the actors (cf. Fig. 1). Sound was recorded using a separate microphone at each of the three cameras, plus an additional microphone positioned over the left ear of the actor, providing a separate speech recording with a constant distance to the actor's mouth.

[1] Definitions and scenarios are currently available only in French and can be obtained on request to the first author.



a b c

Figure 1: still frames illustrating the 3 camera angles used in the video recordings

Video and audio data was segmented on the level of single sentences. Recordings containing the two standard sentences (pseudo-speech) and the sustained vowel, as well as improvised sentences (in French) were extracted and saved in separate files. Over 7'300 such sequences, among them about 5'000 containing the pseudo-linguistic sentences and the sustained vowel, were extracted from the original interactions. However the corpus will be reduced to less than 2'000 sequences by removing redundant sequences; 2-3 repetitions will be selected for every expected condition based on expert ratings.

The selected sequences will be thoroughly documented in terms of lay ratings of emotional expressivity, "naturalness" of expressivity, and accuracy of emotional communication. The audio recordings are currently automatically segmented, i.e. phoneme and syllable boundaries are determined, using phonetic alignment methods.

RESEARCH PLANS AND POTENTIAL APPLICATIONS

The uniform structure of the recordings makes this corpus especially well suited for systematic quantitative analysis and comparisons across emotions. This holds true for both the acoustic as well as the visual aspects of the data. The first results of the phonetic alignment and segmentation of the standard sentences are promising. The prospective availability of fine-grained phonetic segmentations will allow for phonetic studies that go well beyond the measurement of global prosodic parameters. For example durational effects can be measured at the level of single phonemes. In addition an exact phonetic transcription makes investigations on pausing and articulatory precision possible. Special attention will be paid to the investigation of voice quality parameters. The availability of robust methods for measuring these parameters is still an issue. But due to its size and diversity the corpus can also be used as either training- or test-data for feature-extraction and classification methods and thus can contribute to the improvement of these methods.

In the visual domain, first trials with hand- and face-tracking are performed. One aim is to prepare the ground for subsequent studies of the effect of emotions on the temporal and spatial properties of hand-gestures, in the sense of Wallbott [7] but with the possibility to use truly quantitative measurements on a much broader data set. Subsets of facial expressions will be manually FACS coded. In a further step, manual FACS annotations could be compared to, and potentially supplemented with, automatic extraction methods applied to facial movements. More importantly, we hope to be able to extract data that will allow comparisons across modalities (voice, gestures, face) for different emotions. One issue in this respect would be to assess the extent of "synchronisation" between modalities for various emotional expressions.

The multimodal nature of the corpus will also allow investigating the effects of the different modalities on recognition accuracy for different emotions in rating tests. It will be possible to evaluate the performance of both human subjects and affect-recognition systems when presented with (unimodal) vocal, facial or gestural/postural expressions in comparison to accuracy for combinations of these modalities.

The selection of states portrayed was essentially driven by research questions and applications stemming from emotion psychology. The recordings should in particular allow to investigate questions related to the processing of emotional faces, voices and gestures in neuro-imaging studies and to develop assessment tools of multimodal emotional sensitivity for normal and clinical populations.

CONCLUSIONS

Acted portrayals can provide very valuable contributions to the study of multimodal expressions of emotion. The emotional states portrayed should be carefully selected and defined according to specific research questions. Nevertheless comparison with expressions recorded under more "spontaneous" conditions remain necessary to qualify the results obtained with portrayals.

Acknowledgements

This research is supported by the Swiss National Science Foundation (FNRS 101411-100367), the EU Network of Excellence HUMAINE (IST 507422) and by the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). OFAI is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology.
This publication reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

REFERENCES

1. Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. Speech Communication, 40, 5-32.
2. Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication, 40, 33-60.

3. Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129, 770-814.
4. Moore, S. (1984). The Stanislavski system. The professional training of an actor. New York: Penguin Books.
5. Russell, J., Bachorowski, J., & Fernandez-Dols, J.-M. (2003). Facial and vocal expressions of emotions. Annual Review of Psychology, 54, 329-349.
6. Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40, 227-256.
7. Wallbott, H. G. (1998). Bodily expression of emotion. European Journal of Social Psychology, 28, 879-896.

Real-life emotions in naturalistic data recorded in a medical call center

Laurence Vidrascu, Laurence Devillers

Computer Science Department,
Spoken Language Group
LIMSI – CNRS B.P. 133
91403 ORSAY CEDEX
FRANCE
{vidrascu, devil}@limsi.fr

ABSTRACT

In the paper we present emotional annotation for a corpus of naturalistic data recorded in a French Medical call center. When studying real-life data, there are few occurrences of full blown emotions but also there are many emotion mixtures. To represent emotion mixtures, an annotation scheme with the possibility to choose two verbal labels per segment was used by 2 expert annotators. A closer study of these mixtures has been carried out, revealing the presence of conflictual valence emotions. Results of the perceptive test show 85% of consensus between expert and naive labelers. When selecting the non-complex part of the annotated corpus, the performances obtained are around 60% of good detection between four emotions for respectively agents and callers.

Keywords

Emotional speech corpora, real-life mixtures of emotion

INTRODUCTION

The majority of emotional speech corpora that have been used in previous research have not been collected in a genuine context with real stimulus events. We argue that “emotion” manifestations are most complex when really felt and not artificially simulated. The term emotion here is used in a general sense to describe emotion related states such as feeling, attitude, mood, interest and also emotions. In the paper we present the annotation of a corpus of naturalistic data recorded in a French Medical call center. An annotation scheme using both dimensions and labels, with the possibility to choose two labels per segment was used by 2 expert annotators for representing emotion mixtures. Emotions are rare events, an observation that was verified in a financial corpus also recorded in a call center that we studied with 11% of the utterances annotated with non-neutral emotion labels. This corpus is richer with 30% of the utterances annotated with non-neutral emotion labels. Still, when studying real-life data, there are few occurrences of full blown emotions but there are many emotion mixtures [1, 2]. A closer study of these mixtures has been carried out, revealing the presence of conflictual valence emotions. Inter-annotation measures and intra-annotation measures were used to assess the reliability of the annotations and study emotion mixtures. A perceptive test was made on a sub-corpus to validate the annotations and assess the difficulty of an emotion detection system working out of context. A comparison of the labels attributed to a sample of 41 utterances from the corpus by experts annotators, by naive users and by the detection system is done. All three provided with different level of knowledge as shown in table 1.

	Expert annotators	Naive subjects	Automatic System
Contextual	X		
Para-linguistic	X	X	X
Lexical	X	X	

Table 1. Different levels of information

Sections 2 and 3 describe the corpus and the adopted annotation protocol. Section 4 relates labels given by naive users through the perceptive test and given by automatic detection with a SVM. Finally, in section 5, the results are compared.

THE CEMO CORPUS

Call centers are interesting solutions for recording people in various natural emotional states since the recordings can be made imperceptibly and they provide real and genuine contexts where emotions are often exacerbate. Our corpus was recorded in a French Medical Emergency call center and is quite rich in emotions. The service center can be reached 24 hours a day, 7 days a week. The aim of this service is to offer medical advice. The agent follows a precise, predefined strategy during the interaction to obtain sufficient details about the situation so as to be able to evaluate the call emergency and to take a decision. In the case of emergency calls, the patients often express stress, pain, and fear of being sick or even real panic. The caller may be the patient or a third person (a family member, friend, colleague, and caregiver) and there are several dialogs involving an agent and two or more callers. 20 hours of speech were thus recorded with 688 dialogs (7 agents and 784 distinct callers with on average 48 turns per dialog). Laughters, coughs, breaths and human noises were written in the transcription. The use of the data carefully respected ethical conventions and agreements ensuring the anonymity of the callers, the privacy of personal information and the non-diffusion of the corpus and annotations. The corpus was annotated with contextual metadata, such as the call origin (from patient, from medical center), the role in the dialog (agent, caller), reason (immediate help, doctor help, and medical information), decision taking etc. Additional information concerning the acoustic quality of the recording (noise, outside/inside, mobile/fixed/radio phone) and the caller information such as sex, age category (child, middle-age, old), accent (French, foreign), defaults in pronunciation/voice quality (nasal twang, lisp, slurred), health/mental state (normal, pathologic, alcohol/drug influenced, hoarse, groggy) are also labeled. Most callers (59.5%) are adult females. The patient makes the call in 30% of the cases. Otherwise, the call is made by a third person that can be more or less related to the patient.

EMOTIONAL ANNOTATION

One of the main challenges we address is the annotation of real-life emotions, requiring the definition of a pertinent and limited set of labels, an appropriate annotation scheme and some reliability measurements.

Annotation Scheme

The audio signal was manually transcribed and further segmented into emotional segments where the annotators felt it was appropriate, so the temporal-grain can be finer than the speaker turn.

Valence	Coarse level (8 classes)	Fine-grained level (20 classes + Neutral)
Negative	Fear	Fear, Anxiety, Stress, Panic, Embarrassment, Dismay
	Anger	Anger, Annoyance, Impatience, ColdAnger, HotAnger
	Sadness	Sadness, Disappointment, Resignation, Despair
	Hurt	Hurt
Negative or Positive	Surprise	Surprise
Positive	Relief	Amusement, relief
	Compassion	Interest, Compassion
Neutral	Neutral	Neutral

Table 2. Emotion classes hierarchy: multi-level of granularity.

20 labels were selected as relevant for the study of the corpus [1].The set of labels is hierarchically organized, from coarse-grained to fine-grained labels in order to deal with the lack of occurrences of fine-grained emotions and to allow for different annotator judgments. The coarse-grained emotion label families and fine-grained labels are given in the table 2. One or two emotional labels could be given per segment. The annotation level used to train emotion detection system can be chosen based on the number of segments available. The repartition of fine labels (5 best classes) only using the emotion with the highest coefficient in the vector [1] is given Table 3.

Caller	**Neu.**	**Anx.**	**Str.**	**Rel.**	**Hur.**	**Oth**
10810	67.6%	17,7%	6.5%	2.7%	1.1%	4.5%
Agent	**Neu.**	**Int.**	**Com.**	**Ann.**	**Sur.**	**Oth**
11207	89.2	6.1%	1.9%	1,7%	0.6%	0.6%

Table 3. Repartition of fine labels (688 dialogues). Other gives the percentage of the 15 other labels. Neu: Neutral, Anx: Anxiety, Str: Stress, Hur: Hurt, Int: Interest, Com:Compassion, Sur: Surprise, Oth: Other.

Validation

The Kappa coefficient was computed for agents (0.35) and clients (0.57) when only considering Major annotation. Yet, we believe measures such as Kappa or Cronbach alpha are poor to assess the quality of an annotation because they can only be applied with one label and don't take into account the fact that there can be several perception for a same occurrence, none of them being "right" or "wrong". Most confusion is between a so-called “neutral state” and an emotional set. Because we argue there can be different perceptions for a same utterance, we considered an annotator as coherent if he chooses the same labels for the same utterance at any time. We have thus adopted a self re-annotation procedure of small sets of dialogs at different time (for instance once a month) in order to judge the intra-annotator coherence over time. About 85% of the utterances are similarly re-annotated. Since segments were labeled by more than one labeler and also since segments could be assigned one or two labels, it was necessary to create a mapping (i.e. to reduce the multiple labels per segment to one label) for the machine learning experiments. Let us consider each annotation as a vector (Major, Minor). The mapping combines the N (Major, Minor) vectors (for N annotators) in an emotion soft vector. Different weights are given to the emotion annotation, one weight to the Major emotions (wM) and one other to the Minor emotions (wm) [1]. About 50% of the corpus was thus labeled as neutral. During the re-annotation procedure, the percentage of the utterances with a neutral label was also computed for each labeler and after the first “training” month, no marked trend towards a diminution of the Emotional labels was found.

Emotion mixtures

Labeler 1 assigned a Minor label for 31 % of the non neutral segments, whereas labeler 2 for only 17 %. Mixed emotions within the same coarse-grained label are noted as **Ambiguous**. A labeler perceiving an emotion between annoyance and hotAnger would label it "Annoyance/Hot Anger". A mixture between two different coarse-grained labels is called **Conflictual** if they don't have the same valence, **Non-conflictual** otherwise. The non-conflictual mixtures can be separated into positive and negative. When perceived with another emotion, the class 'Surprise' doesn't fit into those categories because its valence is not set, which accounts for a class **Surprise**. For analysis purposes, the Conflictual and Unconflictual emotions mixtures are manifestly the most interesting data. It is to be noted (see Figure 1) that both annotators have perceived mixtures in those classes and that they appear in different positions in the dialog (i.e. for Agent and Caller). Agents showed impatience/anxiety mixtures when they identified a high level of emergency and experienced difficulties in dialoguing with the caller (difficulty of understanding non-native persons, social differences, physical condition). For the callers, the most frequent mixtures involved relief/anxiety, positive/stress which at the first view seem impossible to obtain. Such conflicting emotions are often observed near the end of the dialog, when the person knows that help is coming, but still remains fearful about his condition. Evidence suggests that such a perception is possible, because the two emotions are expressed at different levels, one linguistic and contextual and the other paralinguistic.

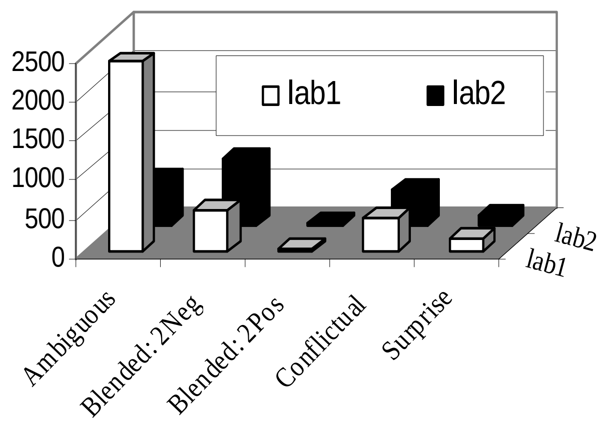


Figure 1. Repartition of the mixed emotions for each labeler (Lab1 and Lab2); Blended: 2Pos means that the two labels are chosen from two different positive labels ('Amusement', 'Relief, 'Compassion/Interest'); Blended: 2Neg means that the two labels are chosen from two different negative coarse grained labels ('Fear', 'Anger', 'Sadness', 'Hurt').

PERCEPTIVE TEST

Protocol

There were several reasons for doing the test. The first one was to validate our annotation scheme. We wanted to substantiate that our set of labels was appropriate for our corpus and that the distinction between fine grained labels such as compassion and interest or annoyance and impatience was relevant, despite their confusion in some cases. We also wanted to answer the criticism that 2 annotators might not be enough to annotate the corpus reliably. The other goal of the test was to appraise the presence of emotion mixtures and see if the lack of context would hinder the perception of conflictual segments. 41 segments were selected including 14 “simple” segments (annotated with one emotion by both annotators), 11 “non conflictual” segments, 13 conflictual segment and 3 complex segments on which the 2 annotators didn’t agree. A typical case of a conflictual emotion for an agent is to feel both annoyance and compassion towards a caller (or try to hide his irritation with a compassionate voice). An example in the sample used for the test is "I understand you, madam, but I have no miracle solution". For the client, recurring examples were to feel worry coupled with relief from knowing help is on its way or embarrassment and self amusement like a student babysitting who when asked for the apartment number answers with an embarrassed laugh "Let me take my notebook because I can't even remember the number here". The stimuli were given without context but could be replayed indefinitely. The test was run with 43 subjects: 33 French native people (13F/20H) and 10 non native French speakers. The subjects were first asked to rate the valence of the stimulus (from very negative to very positive) on a scale from -3 to 3. Then they had to choose a label for the emotion they perceived in the list of 20 labels + neuter. In the case when a second emotion was perceived, they had to choose it from the same list, and clarify if the 2 emotions were perceived simultaneously or sequentially. Finally, they were to write the name of the emotion they perceived if it wasn’t in the list. Because of the absence of context and of the large number of labels, this task felt very difficult, especially for the non native subjects. Most of them were unable to specify mixtures of labels. Thus we only considered native subjects when studying emotion mixtures.

Evaluation Results per subject

Without the context, every native French subject perceived some occurrences of emotion mixtures and all but 2 subjects among them perceived mixtures of a positive and negative emotion.

Annotated as->	**Simple/ Ambiguous**	**Non conflictual**	**Conflictual**
Simple (14 seg)	87%	7%	6%
Non conflictual (11 seg)	76%	19%	5%
Conflictual (13 seg)	71%	10%	18%

Table 4. Percentage of simple and complex emotions

Yet on average a subject could perceive 9 segments with emotion mixtures (with some on “simple” segments), which is very low. Table 4 shows for each subset (simple emotion, non conflictual, conflictual) the percentage given by naive users. For 46% of the conflictual sample, people were able to perceive emotion mixtures (mainly conflictual ones). In the other hand, there were still 22%

which were judged as complex when annotated as simple. In this study, women perceived more conflictual mixtures than men. These poor results show the difficulty of the perception of these samples without context.

Evaluation Results per vector

Even when subjects individually chose one label, the vector combining the annotations of all 45 subjects appears to correspond to the vector of the 2 expert labelers. Indeed, when comparing the two highest coefficients of the vectors for expert annotators and naive annotators, there is an agreement of 85% between the two annotations. 70% of the complex emotions were detected (9 segments out of the 11 non conflictual and 9 out of the 13 conflictual have the same 2 coarse emotions). Errors often involve relief that out of context is labeled as fear. The cases where experts and naive annotators disagreed were accounted for by the context, such as for instance "Oh poor you, oh poor you, it's already been 2 years" that was perceived as compassion but was in fact was an answer to the question “how long have you had trouble breathing?” and in the context was clearly perceived as fear and sadness by both expert annotators. For the utterances on which the annotators did not agree on, no consensus was found among naive annotators as well.

Evaluation Results with an SVM classifier

In order to compare the annotations with context, lexical and prosodic content (expert annotators) to those with lexical and prosodic content (naive subjects) and to the only prosodic ones, the first step was to create a model for emotion detection with the classes adapted to the perceptive test. For the experiments about 100 features were extracted including prosodic and spectral features obtained with the Praat program [3] and lexical features extracted from the dialogs transcription such as non-verbal events or hesitations. An exhaustive list of the parameters can be found in [4]. We used the weka software [5] with SVM classifiers for training the data. Because previous experiments had shown that agents and callers have different expression of emotion (see table 4) [1, 4], we made one model for each role.

Callers	**Total**	**Fea**	**Sad**	**Ang**	**Rel**
#Utterances	640	384	100	49	107
% rec.	**59,1**	59	65	39	62
Agents	**Total**	**Neu**	**Ang**	**Comp**	**Surp**
#Utterances	810	346	155	285	24
% rec.	**54,9**	60	55	52	33

Table 5. Detection Results with a SVM classifier and 4 classes for Agents and Callers: Ang: Anger, Ang: Anger, Fea: Fear, Neu: Neutral, Rel: Relief, Sad: Sadness

We only used emotion labels (the macro-classes) which are in the test. The Agent model was made with the 4 macro-classes Neutral, Anger, Surprise, Compassion (which includes Interest) and the caller’s model with the classes Fear (which included Stress), Anger, Sadness and Relief. We can pinpoint that Anger is not quite the same when expressed by callers or by Agents. Non complex utterances that both annotators agreed on were used to train the models with about 300 utterances per emotion. The train set and test set contained utterances from different speakers. The utterances selected for the perceptive test were part of the test sets. Table 5 shows the emotion detection results using SVM classifiers. Predictions made on the segments from the perceptive test were compared to the expert and naive annotations. The percentages of good rating are given in table 6 below.

e=n=w	61 %
e=n	85 %
e=w	66 %

Table 6: percentage of agreement only considering the highest coefficient of the annotation vectors, e: expert annotators, n: naive annotators, w: automatic detection

CONCLUSION

This paper focuses on real-life emotions and shows the complexity of natural emotional behavior expressed in dialogs on a medical call center. Our study of this speech corpus reveals the presence of mixtures of emotions with conflictual or non-conflictual valences. Results of the perceptive test show 85% of consensus between expert and naive labelers. We have shown the meaningfulness of the context for annotating emotion. When selecting non-complex part of the annotated corpus, the performances obtained are around 60% of good detection between four emotions for agent and callers when only using paralinguistic information. Further studies will be to combine paralinguistic, lexical and contextual information to improve real-life emotion detection.

Acknowledgements

This research is partly supported by the EU Network of Excellence HUMAINE. This work is conducted in the framework of a convention between the APHP France and the LIMSI-CNRS.

REFERENCES

1. Devillers L., Vidrascu L. & Lamel L., Challenges in real-life emotion annotation and machine learning based detection, Journal of Neural Networks 2005, 18/4, numéro special "Emotion and Brain".
2. Douglas-Cowie, E., Devillers, L., Martin, J-C., Cowie, R., Savvidou, S., Abrilian, S., Cox C., Multimodal Databases of Everyday Emotion: Facing up to Complexity, Interspeech 2005.
3. Boersma P., Weenink, D., Praat: doing phonetics by computer, from http://www.praat.org/ (2005)
4. Vidrascu L., Devillers, L., Real-life Emotions Representation and Detection in Call Centers, ACII'2005.
5. Witten, I.H., Franck, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S.J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. Proc ANNES'99 International Workshop: Emerging Engineering and Connectionnist-Based Information Systems, 192-196.

Collection and evaluation of an emotional speech corpus using event recollection

Noam Amir

Dept. of Communication Disorders
Tel Aviv University
noama@post.tau.ac.il

Samuel Ron

Center of Mental Health
Tel Aviv University
sam69ron@yahoo.com

ABSTRACT

In the present study we discuss a method of emotional speech elicitation, event recollection, and methods used to validate a corpus based on this technique. Participants were seated in a comfortable environment and asked to recollect several emotional episodes in their past, trying to feel as they did when these took place. They were asked to experience the following emotions separately: anger, fear, joy, sadness and disgust. Data was collected from 136 subjects, resulting in a corpus of 1630 minutes of speech. We describe various methods used to evaluate the emotional content of the corpus and discuss the problems that crop up in such a process.

Keywords

Emotions, speech, listening tests.

INTRODUCTION

Collection and evaluation of speech corpora reflecting authentic expressions of emotions is necessarily the basis for any research on speech and emotion. Both collection and evaluation are difficult tasks each in their own right. Many approaches to these problems have been attempted, with their associated compromises as to recording quality, textual content, speaker variation, variation in emotional content, variation in evaluation, and other factors.

From the collection side, one approach that has been used extensively is to record acted emotions, from professional or non-professional actors, as done by Banse and Scherer and others [1, 2, 3]. Whether the manner in which emotion is expressed by actors is the same as in spontaneous speech has been discussed in various papers [4, 5, 6], though there is no wide agreement between researchers on this subject.

In recent years there appears to be a trend toward studies based on more naturally occurring speech, from various settings such as television talk shows [4], dialogs recorded during the performance of a given task [6, 8], televised "reality shows", event recollection [5], speech taken from call-center data [9, 10], speech elicited by inducing emotions in non-acting subjects, etc. To different extents, all the methods that do not involve actual acting suffer from at least some of the following difficulties:

1. Confidentiality and/or copyright issues
2. Sparsity or small variety in emotions expressed
3. Difficulties in reproducing results
4. Lack of control over the expressed emotions
5. Lack of extreme emotion or inhibitions in expressing them

Amongst the different methods above, the event recollection method we describe here strikes a reasonable compromise with respect to these issues. In this method subjects are asked to recall and re-experience emotional episodes, and then describe them as they are experiencing them. This method can be applied in a controlled laboratory setting to elicitation from any selected subjects (both pathological and normal), and it can be used to elicit a large range of emotions. A major drawback of this method is that subjects often tend to lapse into "narrative" mode, where they are describing an event but not experiencing an emotion. Such epochs can be examined through two complementary methods: listening tests and physiological measurements.

Employment of such complementary methods is extremely interesting, though it is most definitely not a panacea. In fact, it can often raise more questions, when recording different instances in which there is a clear manifestation of emotion as evaluated through 1) speech only, 2) physiology only, or 3) both speech and physiology. Thus, as each method has its advantages (and pitfalls), it seems that the research performed using one method of analysis does not necessarily apply to the other method of analysis.

The issue of listening tests, which we concentrated on here, adds a further level of complication. One side of the problem is purely practical: subjecting large corpora to extensive, detailed listening tests is extremely time consuming, fatiguing to the listeners, and can be quite expensive. Obtaining good agreement between listeners can also be difficult, depending on the variance in age, background and personal tendencies of the listeners. The other, more theoretical aspect is the question of emotional labeling. As in emotional elicitation, there is no widely accepted procedure for labeling emotions. Much of the initial work in this domain cited several "basic" emotions such as anger, joy, sadness, fear, disgust, surprise and others, though in recent years there has been a trend towards other labeling schemes. These have been either towards more refined categorical labeling, or towards dimensional psychological scales that are not emotions in themselves, such as active/passive, positive/negative, empowered/submissive. It is interesting to note that even though the elicitation scheme rests on the notion of there being a set of "basic emotions", the labeling scheme does not require the use of the same method.

The purpose of the present study was to accumulate a large corpus of emotional speech, which would serve as a basis for extensive acoustic analysis. As mentioned above, the emotion recollection method was chosen for several reasons, and after the recordings were carried out we conducted a listening test in two phases, in order to verify its emotional content. In the first phase a small number of judges listened to large parts of the corpus, giving more general judgments over stretches of speech of approximately one minute in length. In the second phase, 126 small segments of approximately 10 seconds each were extracted from the corpus and subjected to a more detailed listening test, in order to evaluate them more closely.

The results we present here are related to the agreement between judges over the main and sub-corpora. This comparison brings to light several interesting questions as to the influence of textual content, vs. the subjective impression of the judges as to how this content influenced their decision. It reflects how easily emotional speech can become ambiguous when utterances are taken out of context, and shows that the outcome of listening tests, even when carefully carried out, can be influenced by the way in which they are conducted. Analysis of physiological measurements, which were also collected, will be addressed in a separate paper.

ELICITATION METHOD

The participants consisted of volunteer university students, half men and half women. Mean age was 24 years (range 21 to 28). Individuals undergoing neurological or psychiatric treatment or who took medication three days before the study were excluded. Subjects using illicit drugs or alcohol were also excluded from the study.

Subjects were seated in a comfortable environment and asked to recollect several emotional episodes in their past, trying to feel as they did when these took place. For imagery induction, subjects were asked to identify five emotional situations from their own lives and re-experience the feeling that corresponded to each of the following five emotions: anger, fear, joy, sadness and disgust. After concentrating on each experience they were asked to describe it for approximately a minute and a half, during which their speech was recorded. The first 30 seconds were set-aside for "mental setting", so that only the last 60 seconds were retained for analysis. In addition, neutral speech was also recorded. As we are aware that visual percept can influence the recalled emotion, we repeated each trial once with eyes open and once with eyes closed. Thus, the experiment produced 12 recordings for each subject. The data was collected from 136 subjects, resulting in a corpus of 1630 minutes of speech

LISTENING TEST #1

A subset of the corpus, incorporating 30 speakers, was subject to a comprehensive listening test by 6 judges, with the objective of verifying the reliability of the corpus. The listening test was conducted automatically through a simple graphic user interface. Listeners heard each one-minute recording in its entirety, after which they were required to answer several questions before passing on to the next recording:

1. **Emotional content**: listeners assigned a percentage for each of the five emotions, neutral or other, so that the sum was required to be 100.
2. **Credibility**: listeners evaluated credibility of emotional expression on a five point scale from high to not passable
3. **Recognizability**: listeners evaluated recognizability of emotional expression on a five point scale from very high to unrecognizable.
4. **Involvement**: the listeners evaluated whether during the recollection episode the speaker was involved in the description or was only "reporting" as an "observer" vis-a vis the event. This was graded on a five point scale from "all the time" to "not at all".
5. "**Switching**": listeners estimated the number of times that the subject "switched" between participant and observer states, given the possibility of 0, 1, 2-3, more than 3.

Each listening test took about 10 non-consecutive hours. Clearly such a procedure is time consuming and expensive, Thus we were able to conduct it for only 6 judges. The results of such a test still leave a certain amount of uncertainty as to where the more emotional parts reside in each recording. We thus treated this test as a screening procedure, which served as the basis for a further, more detailed listening test, on a reduced amount of data.

LISTENING TEST #2

The objective of the second test was to cull a smaller and hopefully more representative corpus from the larger one, which could fulfill several objectives. A smaller corpus would be far easier to subject to more extensive listening tests, to analyze acoustically in detail, it could be expected to contain more distinctly expressed emotions, and would have a more neutral textual component. The latter requirement stemmed from the fact that over a minute of speech the listeners could quite easily discern the emotion being expressed almost solely from the textual content.

In order to prepare the reduced database, two research assistants scanned the corpus with the objective of obtaining 20 representative extracts of about 10 seconds each for each emotion, which they felt to be authentic expressions of the said emotions. This sub-corpus was then used in a second listening test presented to 19 judges.

In this test, the judges were asked the following questions:

1. "What is the expressed emotion?" (7 point forced choice – 5 emotions, neutral or other)
2. "Strength of the emotion" (weak, medium, strong)
3. "Was the speaker truly experiencing the said emotion?" (yes/no)
4. "Did the textual content influence your decision?" (yes/no)
5. "Did you deliberate between the chosen emotion and another one?" (yes/no)
6. "If yes – which?" (7 point forced choice – 5 emotions, neutral or other)

This listening test took about an hour per subject, over two half hour sessions.

RESULTS

The results of both listening tests were subjected to extensive statistical analysis. For lack of space we will quote the main results, and discuss some questions which arise from these and also from several anecdotal results.

First listening test

Overall ratings of emotional content suggest that on average, listeners achieved high levels of correct identification. Table 1 shows the average ratings of emotional content for each of the emotions, when the stimuli were of the corresponding emotion. Results are presented for male and female speakers separately. Neutral received the highest mean ratings, while other emotions received moderately high ratings. Interestingly, the average ratings for female speakers are consistently higher than for male speakers.

One must keep in mind that over a one minute stretch of emotional speech, numerous cues as to the emotion being described , both prosodic and others, are available to the listener, even if he or she is in a more narrative mode.

This means that high rates of correct evaluation are to be expected, and do not present sufficient proof that the speakers are actually experiencing the said emotions, or expressing them in the paralinguistic features of their speech. Therefore, in order to obtain a general direction as to where to look for the most authentic expression of emotion possible, we later examined the judgments of credibility, in addition to those of emotional content. These are presented in table 2.

Emotion	Male speakers	Female speakers
Anger	74.3(22.6)	86.5(16.4)
Joy	71.6(22.6)	80.1(22.9)
Sadness	76.0(22.2)	81.7(22.2)
Fear	69.1(25.3)	81.5(21.3)
Disgust	77.9(22.6)	81.3(25.0)
Neutral	90.3(17.4)	93.5(15.7)

Table 1. Mean and STD of emotional content ratings, when the stimuli were of the corresponding emotion.

Emotion	Male speakers	Female speakers
Anger	1.84(0.92)	1.70(0.98)
Joy	1.92(0.96)	1.93(1.14)
Sadness	1.64(0.78)	1.76(0.94)
Fear	1.90(0.97)	1.90(1.08)
Disgust	1.79(0.90)	1.87(1.01)
Neutral	1.29(0.54)	1.59(1.20)

Table 2. Mean and STD of credibility ratings (1-very high, 5- not passable)

Though the average ratings are quite high, the standard deviations are also high, indicating that this measure can be useful in differentiating between passages that were more truly emotional in the judges' opinions. These were taken into account when extracting the shorter excerpts for the second listening test.

Second listening test

The responses in this test were analyzed in detail, examining the recognition rates and its interactions with other responses. For lack of space we cannot quote many of these. The main recognition results are presented here: in order to analyze the number of correctly identified excerpts, it was necessary to determine a threshold for deciding whether a specific excerpt was judged correctly. We examined the results with 3 different threshold levels: 3, 7 and 11. The results appear in figure 1. Chance level is at about 3 correct judgments per stimuli, and 11 is well over 50% correct. Results for thresholds of 7 and 11 are similar.

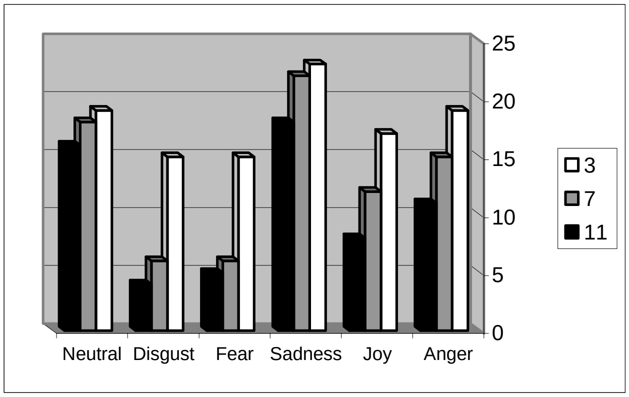


Figure 1. Number of excerpts identified correctly using different thresholds

Some interesting observations stem from more anecdotal results also, mainly from the point of view of the influence of textual content on the judge's decisions. This is quite hard to determine objectively, but several impressions we obtained are worth mentioning.

For example, the transcription of one of the excerpts is: "the whole family … and his children". The prosody is so overwhelmingly sad, that 9 listeners out of 19 said that the text influenced their judgment. Clearly the text is completely neutral, and could just as well be a description of a joyous occasion such as a family reunion.

The same observation holds for another excerpt, transcribed as: "and I returned from a trip outside with my niece, and I came in and my mother was sitting inside". Once again, the text is neutral in our judgment, yet 13 of the listeners said that the text influenced their judgment.

Content can be quite misleading on the other hand, especially when the prosody is ambiguous. In an excerpt transcribed as: "you want to shout and you shout till it comes out as… till you all get to …" Evidently this excerpt contains some ominous overtones, since 9 listeners judged the emotion as fear, whereas it is taken from the description of a soccer game in which the speaker's team was winning. Furthermore, 13 of the listeners reported that the text did in fact influence their judgment. It is interesting to note that fear is often regarded as an emotion where the speaker tends to be passive, though this is clearly not the case in this excerpt.

DISCUSSION

The results of the first listening test indicate that correct identification of the correct emotions, as reflected in their content ratings, are high. This is probably due to the mixture of textual/non-textual content, though credibility ratings can give some finer indications as to which is more dominant.

The second listening test shows that over shorter excerpts the paralinguistic features tend to become dominant and the influence of the textual content is reduced, though in fact the listeners themselves are poor judges of this. In this case, the results became far less uniform. Here, sadness was most easily identified. Disgust, on the other hand, was often confused with sadness or fear, and fear also was often confused with sadness. Anger, joy and neutral were also mostly identified. In addition, the degree of identification was correlated with the degree of emotion, which is to be expected.

REFERENCES

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

2.Yang, L. and Yunxin, Z., Recognizing emotions in speech using short tem and long term features, Proceedings of ICSLP 98, Sydney

3.Dellaert, F., Polzin, T, Waibel, A., Recognizing emotion in speech, Proceedings of ICSLP 96

4.Douglas-Cowie, E., Cowie, R., Schroder, M., A new emotion database: consideration, sources and scope, ISCA workshop on speech and emotion, Belfast 2000

5.Amir, N., and Ron, S. Towards an automatic classification of emotion in speech, Proceedings of ICSLP 98, Sydney

6.Kehrein, R., Prosodie und Emotionen, Tuebingen Neimeyer 2002

7.Cowie, R.,Describing the emotional states expressed in speech, ISCA workshop on speech and emotion, Belfast 2000

8.Fernandez R., Picard R. W. , (2000), Modeling drivers' speech under stress, ISCA workshop on speech and emotion, Belfast 2000

9.Ang, J., Dhillon R., Krupski A., Shriberg E., Stolcke A., "Prosody based automatic detection of annoyance and frustration in human-computer dialog", ICSLP 2002.

10.Devillers L., Vasilescu I., Lamel L., "Annotation and detection of emotion in a task-oriented human-human dialog corpus", ISLE workshop, 2002

MEED: the challenge towards a Multimodal Ecological Emotion Database

M.R.Ciceri[1], S. Balzarotti[1], F. Beverina[2], F. Manzoni[1], L. Piccini[3]

[1]Lab of Communication Psychology, Catholic University of Sacred Heart, Italy
[2] STMicroelectronics, Advanced Research Group, Italy
[3]Bioengineering Department, University Politecnico di Milano, Italy

ABSTRACT
This study is aimed at giving a contribution in the development of emotional corpora and databases, trying to face the theorical and methodological issues that recent research has brought to evidence. It is in fact essential to understand how to collect and organize all the different indexes and signals related to the same emotional events. In particular, concerning ecological validity, film segments and videogames were used to achieve emotional induction. Dealing with multimodality, we tried to combine physiological signals such as Galvanic Skin Response and Electrocardiogram on one side, and non verbal signals such as facial expressions, postural movements and vocal behaviour on the other. Finally, a dimensional system referring to appraisal theories was used to label data. Among the available emotional databases, this one has the unicity to have audio-visual signals, synchronized with physiological signals and a set of eliciting events.

Keywords
Emotion; Ecological Database; Multimodality.

INTRODUCTION
In the last decades, different theories on emotion and various definitions have been tried [10] [21] [22] [20] [16]. Although there is no consensus among authors on what the structure of emotion should be [20], it seems possible to find a common emphasis on the multi-componential nature of the emotional phenomenon [22]. Several groups of research are now working at the development of emotional corpora and databases, thus opening theorical, methodological and technical issues concerning how to collect and organize all different emotional indexes, such as physiological signals, facial movements, gesture and vocal expressions. A first concern is multimodality. Many of the existing and wide used databases have started mainly focusing on one emotional expression system [8][17][26]. Anyway, in the last years, there is a growing interest in multimodal databases [2] [11] [18]. A second main issue may concern naturalism, that is the importance of considering emotions as they occur in everyday life. Many emotional corpora are acted or posed, so that subjects are directly asked to simulate different kinds of emotion [14] [26] [3]. Trying to ensure both standardization and high quality measures on one side - which are crucial for research purposes - and ecological issues on the other still seems not easy: for that reason, recent research has been concerned with problems about acted data and various methods have been explored moving towards more ecological collections [2] [11] [18]. Finally, attention should be paid to which kind of emotion theories are needed: this seems a key issue as it involves different choices about the type/number of emotions considered and the labelling techniques selected. For example, a categorical approach considers the expression of a few archetypal or discrete emotions [14], while a dimensional one aims at the analysis of streams of different behavioural responses in dynamical emotional episodes reflecting underlying appraisal processes [13].

EMOTIONAL INDUCTION
In this study, we aim at giving a contribution in this area as an initial attempt in the development of an emotional corpus. In particular, we tried to face:
(1) Ecological validity: we decided not to collect "acted" expressions, trying to include naturally occurring samples of emotions. To this purpose, two different kinds of stimuli were used to achieve emotional induction:
(a) In a first condition movie segments were presented. According to literature, movie narrations can arouse emotions similar to those of everyday life [7] pertaining to the category of "witness emotion" [25]. Gross and Levenson [12] claim that movies can be used as experimental stimuli able to elicit emotional responses in all different multimodal systems of emotion. Other authors suggest that besides its narrative structure, cinema distinctive features such as shots, camera movements and image speed on screen can deeply influence the emotional response in viewers [9] [23]. In this condition, 36 (18 male and 18 female) subjects were video recorded while they were watching samples from three different movies (Dead Poet Society, Thelma and Louise, The last kiss). To 18 subjects emotional scenes (30-40 seconds length) extracted from longer sequences (emotional expression only) were shown. The other 18 watched extended stimuli (lasting 11 minutes) where the same scenes were included in a

clear narrative structure [24] (antecedent + emotional expression).
(b) In a second experimental setup subjects were video recorded while playing with videogames. Experimental studies on emotion have suggested the use of computer games [13] [15], because they are better controllable than real interactions and offer an experimental situation where the subject is involved not just as a simple observer but has to act and react [15] [27]. 24 (12 male and 12 female) undergraduate subjects were asked to play with four different kinds of computer games especially developed to elicit specific emotional reactions. Systematic manipulation of the appraisal dimensions was used through the selection of types of game events that were assumed to produce specific appraisals [27]. Specifically, game events were supposed to support four emotional evaluation checks (novelty, hedonic value, goal conduciveness, coping).
(2) Multimodality: in this paper we tried to combine different signals simultaneously: physiological signals such as GSR (Galvanic Skin Response) and EGC (Electrocardiogram) on one side, and non verbal signals such as gaze, postural, facial movements and vocal behaviour on the other.

EXPERIMENTAL SETTING

Different kinds of devices were used to record the subject's behaviour: two kinds of web cameras - Philips toUcam pro II (640x480 - 25 frames/ sec), STM web-cam (320x240 – 25 frames/sec)-; the BIOPAC™ System (Biopac Inc., Goleta, CA, USA) [4] to record physiological signals; a high quality microphone to record vocal reports. All instruments were synchronized through the use of an external trigger, which drove the BIOPAC system. Such trigger was generated by the main computer, which stored also the video and audio information. The biological signals (ECG, GSR, Respiration, Skin Temperature) were acquired with standard non invasive techniques and the data were collected from a second PC connected to the BIOPAC.

PROCEDURE

On the arrival, participants were seated in a well-lit room. They were informed that they would be video recorded and their prior consent was asked for the treatment of personal data only for research purposes.

Figure 1. Experimental setting

In the first condition (a) subjects were asked to watch different movie samples and, at the end of the experimental session, to complete a self-report questionnaire and to choose from a given list of terms (based on the Circumplex model, [19]) in order to judge their emotional experience. In the second condition (b) subjects were asked to use the computer where an avatar (Baldi, [1]) guided them across different kinds of computer games. All sessions started with 2 minutes of free exploration of a number of web pages regarding descriptions of university courses assumed to be emotionally neutral. This was done in order to have a profile of the physiological baseline for each subject.

DATA EXTRACTION CRITERIA

This work presents an initial contribution to develop an organized corpus of emotional signals where different types of data are provided for each subject. The attempt is thus to offer the chance to simultaneously and dynamically examine what is happening in different response systems during the flow of an emotional episode:

- Video and audio (audio is available only in the videogame condition); the camera is pointing at the subject's face and shoulders (frontal view);
- Corresponding film segment extracted and matched to the subject's video for the first kind of experiment; for the computer game experiment we recorded the event which subjects undergo;
- Physiological signals (ECG, Respiration, GSR, Skin Temperature);
- Behavioural coding: at a first level of analysis, all videos were coded *frame by frame* (25 fps) using The Observer 5.0 Software (NOLDUS, Wageningen, The Netherlands) by three independent coders. Reliability was calculated (mean value Kohen's Kappa=.89)
- This analysis had previously required the selection of behavioural units to be extracted (BCS, Behaviour Coding system) [5]. Four categories were considered:
    - *Facial movements*: the fundamental muscle movements that comprise Facial Action Coding System [6] were selected. We considered action units relating to upper face and lower face (20 AU and 10 AD).
    - *Gaze direction:* we considered if the subject looks at the screen, at the keyboard, around, etc.
    - *Posture*: behavioural units of moving near to /far from the screen were considered.
    - *Vocal behaviour:* it was recorded when the subject speaks (verbal speech) or uses other kind of vocalizations: non linguistic sounds, reflex sounds, exclamations and interjections, laughing, panting.
- At a second level of analysis, data were labelled using the Multidimensional Emotional Appraisal

Semantic space (MEAS, [5]), that is a dimensional system to detect the user's emotional state. MEAS, derived from SECs [21], analyses the intersection of four dimensional axes: besides the usually considered dimension of activation (arousal) and evaluation (or hedonic value), it considers the two additional axes of novelty and coping. Three axes (novelty, hedonic value, coping) were scored on a 5 point rating scale from -2 to +2 by three judges according to rules that combined non verbal signals and computer events (context). Inter-judge agreement was calculated (Kohen's kappa =.81). Information on the forth dimension concerning arousal were derived from features extracted from the ECG and the GSR signals.

- Antecedent stimuli analysis (cinematographic language, camera movements, game events);
- Self-report measure.

RESULTS

In figure 2 we present a selected sequence to offer an example and show an initial and still limited attempt to investigate how multimodal sources of information can be combined and labelled: physiological data (HR and GSR), non verbal behaviour, task events, subjects' performance and finally the emotional axes which should represent the constantly changing emotional appraisal of the situation. From a qualitative analysis of this extracted data, it is possible to observe: (1) several similarities are present between the most pronounced variations in physiological signals and changes in the dimensional axes; in particular, the novelty dimension is the mostly related to the bio-signal changes; (2) we need to rely on changes in multiple modalities to enhance the interpretation level of the emotional experience; (3) both variations in physiological and non verbal signals seem to be congruent to the task events; (4) non verbal behavioural units (especially facial movements) and physiological signals seem to require different temporal units of analysis: a 25 fps resolution allows to observe the onset, offset and intensity micro-variations of action units, while a shorter one is required in order to extract useful information by physiological variations, although these latter are more reliable observing a longer time-scale, compatible with the human autonomic nervous system dynamics.

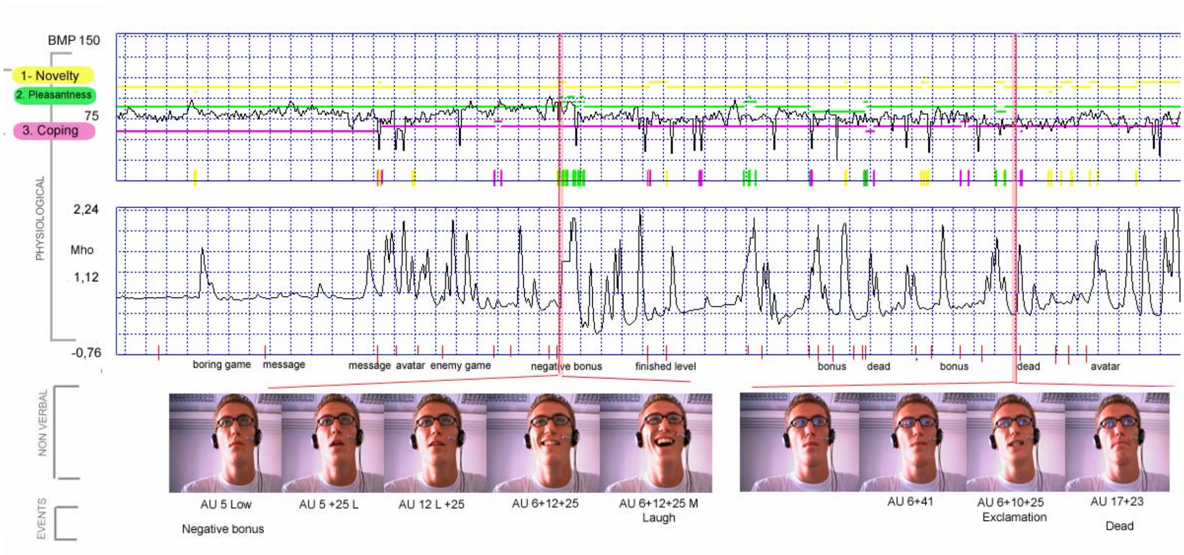


Figure 2. 1) HR and EDA corresponding to a temporal section of 16 minutes; 2) dimensional axes of novelty, pleasantness and coping on a rating scale ranging from -2 to +2; 3) two frame sequences from video intervals of 3.18 and 1.80 sec showing non verbal units in correspondence of 4) a specific event (negative bonus) and performance (dead).

LIMITS

As said above, this study represents a contribution in developing an emotional corpus, and it is open to many problems and limits we hope to solve in the future. Here we list some of the possible issues and challenges.

Ecological vs. reliable data?

In the collection of emotional signals, we chose to collect data as much representative of everyday life as possible in a semi laboratory situation, hence trying to combine ecological validity with experimental control. Anyway this choice was obviously open to methodological and technical problems, which inevitably influenced the recording quality of our data. In particular, facial recordings were characterized by arbitrary poses, head and posture movements, hands sometimes covering the face, people exiting from camera field of vision, etc. As the expressions are not acted, but spontaneously exhibited, data does not include the same fixed configuration for all subjects, but are characterized by more variability among them.

Exploiting physiological signals

Movements also affected physiological signals reliability. An unobtrusive, ecological measure exploited through standard systems (like BIOPAC) increases the possibility of artefacts and of collecting noise on biosignals. This implies the use of more demanding processing in order to extract reliable information. It is possible to use specific algorithm dedicated at removing or identifying such artefacts.

Emotional labelling: which rules?

The MEAS system we used in labelling our data refers to a dimensional approach [13] [22]: the main goal is to analyse less intense or less prototypical emotions and the emotion related states (interest, boredom, satisfaction, etc.) and not only the representation of discrete emotions. If this - we believe - is the strength of the system, weaknesses are also present.

First of all, our attempt was to define rules able to link single or multiple observable emotional responses (e.g. facial action units) to underlying appraisal dimensions. These rules were derived in part from literature (for example, eyebrow raising as indicator of novelty) and in part from the multimodal behavioral analysis we reported above, hence involving a both top down and bottom up process. For this reason the rules are strictly linked to our data and will require further validation if we want them to be generalized to other contexts. Generalization seems to be limited by another consideration: we decided to include game events and performance (easy vs. difficult question, wrong vs. correct answer, next game level, etc.). This decision was anyway supported by an embodied approach to emotion: the emotional meaning of a behavioural expression can be determined only within the temporal and situational context in which it occurs.

ACKNOWLEGEMENTS

This study is part of a wider research project and we would like to thank all the people who were involved in its realization; a special thank to Prof. Bruno Apolloni of the DSI Department, State University of Milan.

CONTACT INFORMATION

Address correspondence to: 1) Lab of Communication Psychology, Catholic University of the Sacred Heart, Largo Gemelli, 1, 20123 Milano, Italy; Tel: +39-02-72343734, Fax (Italy): +39-02-72342800; e-mail: maria.ciceri@unicatt.it; 2) STMicroelectronics, Advanced Research Group, via Olivetti 2 Agrate Br. Italy, fabrizio.beverina@st.com; 3) Bioengineering Department, University Politecnico di Milano, Via Garofalo, 39, 20133 Milano Italy; Tel: +39-0223999021 Fax (Italy): +39-02-23999000; e-mail:luca.piccini@polimi.it.

REFERENCES

1.Baldi CSLU ToolKit. Website http://cslu.cse.ogi.edu/toolkit/.

2.Belfast Naturalistic Database, Douglas-Cowie et al., 2000; 2003.

3.Berlin Database of Emotional Speech, Kienast and Sendlmeier, 2000.

4.Biopac Systems inc (accessed mar 2004). http://www.biopac.com.

5.Ciceri, R., Balzarotti, S., Colombo, P., Analysis of the Human Physiological Responses and Multimodal Emotional Signals to an Interactive Computer, In: *Proc. AISB 2005*.

6.Ekman, P., Friesen W.V. and Hager, J., *Facial Action Coding System. The Manual*. Research Nexus Division of Network Information Research Corporation, Salt Lake City, 2002.

7.Eitzen, D. (1999). The Emotional Basis of Film Comedy. In Carl Plantinga and Greg M. Smith (eds.), Passionate Views: Thinking about Film and Emotion, Johns Hopkins University Press, Baltimore.

8.Engberg, I. S., Hansen, A. V. et al. (1997) Design, Recording and Verification of a Danish Emotional Speech Database. *Proc. EuroSpeech*, 1997. Feagin, S.L. (1999). Time and Timing. In Carl Plantinga and Greg M. Smith (eds.), *Passionate Views: Thinking about Film and Emotion*, Johns Hopkins University Press, Baltimore.

9.Frijda, N.H., *The Emotions*. New York: Cambridge University Press, 1986.

10.Geneva lost luggage database, Scherer and Ceschi, 1997; 2000.

11.Gross, J.J., Levenson, R.W. (1995). Emotion Elicitation using Films. Cognition & Emotion, 9, 87-108.

12.Kaiser, S., Wehrle, T., Facial Expressions as Indicators of Appraisal Processes, in K.R. Sherer, A. Schorr, T. Johnstone (Eds.), *Appraisal Theories of Emotion: Theories, Methods, Research*, New York, Oxford University Press, pp 285-300, 2001.

13.Kanade, T., Cohn, J. and Tian, Y., *Comprehensive Database for Facial Expression Analysis*, 2000.

14.Kappas, A., Pecchinenda, A., Don't wait the Monsters to get you: A Videogame Task to manipulate Appraisals in Real Time. *Cognition and Emotion*, 13: 119-124, 1999.

15.Lazarus, R.S., *Emotion and Adaptation*, New York, Oxford University Press, 1991.

16.M2VTS Multimodal Face Database: http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html

17.ORESTEIA Database, McMahon et al. 2003: http://manolito.image.ece.ntua.gr/oresteia/.

*18.*Russell, J.A. (1980). A Circumplex Model of Affect. *Jounal of Personality and Social Psychology*, 39, 1161-1178.

19.Russell, J.A., Core Affect and Psychological Construction of Emotion, *Psychological Review*, 110 (1), pp 145-172, 2003.

20.Scherer, K.R., Studying the Emotion-Antecedent Appraisal Process: An Expert System Approach. *Cognition and Emotion* 7:325-355, 1993.

21.Scherer, K.R., Appraisal Considered as a Process of Multilevel Sequential Checking. In: K. Scherer, A. Schorr, T. Johnstone, (Editors) *Appraisal Processes in Emotion: Theory, Methods, Research. Series in affective science*, London, Oxford University Press, xiv, 478 pp, 2001.

22.Smith, G.M. (1999). *Passionate Views: Film, Cognition and Emotion*. Co-editor with Plantinga, C., Johns Hopkins University Press, Baltimore.

23.Stein, N.L., Glenn, C.G., An Analysis of Story Comprehension in Elementary School Children, in R. Freedle, *New Directions in Discourse Processing*, Norwood, NJ. Ablex Publishing Corporation, 1979

24.Tan E.S. (1996). *Emotion and the Structure of Narrative Film. Film as an Emotion Machine*. Hillsdale, N.J., Erlbaum.

25.The Yale Face Database: http://cvc.yale.edu/projects/yalefaces/yalefaces.html

*26.*Van Reekum, C.M., Johnstone, T., Banse, R., Etter, A., Wehrle, T. and Scherer, K.R., Psycho physiological Responses to Appraisal

27.Dimensions in a Computer Game. *Cognition Emotion*, 18(5): 663-688, Aug 2004.

The Relationship between Acted and Naturalistic Emotional Corpora

Steffi Frigo

School of English
Queen's University, Belfast, BT7 1NN, United Kingdom
schmeff12@hotmail.com

ABSTRACT

With increasing interest in human-machine interaction, new corpora of naturalistic emotional speech are being developed. Little is known, however, about the relationship between acted and naturalistic emotional material. In order to narrow the gap in knowledge, acted and naturalistic emotional corpora are comparatively assessed using multiple layers of analysis to ensure a full and complete picture of the relationship between the two and the reasons behind it. Initial results show strong similarities between some acted and naturalistic material on an instrumental level. This does not appear to be borne out on a close comparative textual analysis of speech from acted and naturalistic corpora.

Keywords

Emotional speech, corpora, acted data, naturalistic data.

INTRODUCTION

With developing interest in human-machine interaction, there is a new impetus to develop emotional speech corpora that are as naturalistic as possible in order to better train emotion-oriented systems [1]. While much of the work that has been done in previous decades has relied largely on acted corpora, these are widely accepted as problematic in many areas and the drive is now to develop naturalistic corpora that transfer more easily to practical contexts. At the same time there are problems associated with the collection of data, whether acted, naturalistic, induced or application-driven.

In this light, one area that deserves greater attention is the relationship between acted and naturalistic corpora. There exist many studies, from as early as the 1930s that have dealt solely with acted emotional speech, and it is necessary to assess their relevance to the current state of research.

The main reason that acted corpora are so commonly used is the relative ease with which they can be compiled, and this has proved attractive to researchers. They have usually been recorded in laboratory settings, using actors or students, and have used written material as their starting point. Actors are simply asked to read or 'act' a given piece with a particular emotion. Many studies, however, do not have sufficient methodological detail for us to be able to make even these assumptions.

Acted corpora have a number of inherent problems stemming from this method of compilation. There is a great difference between acting and reading a passage for recording. There are problems with the varying ability of the actors. There is great difference in the length of speech samples. Much of the material illustrates maximum emotional intensity, and we know less about low intensity emotional speech. Little work has been done on the development of emotion over time. Most studies have focused on a narrow group of instrumental and acoustic measures. Timing, pausing, paralinguistic features and perceptual assessment have often been ignored and there is a need to move away from global measures and return to more linguistically dependant assessment.

The most fundamental area that has escaped scrutiny, however, is how closely acted emotional speech is related to natural emotional speech. Whilst previous studies using acted corpora have their problems, they do present us with a wealth of information the value of which has not been fully realised. In order to assess the relationship between acted and naturalistic corpora, it was important to conduct a study examining the core themes that have emerged from the years of acted emotional speech research and reevaluate them using both an acted and a naturalistic corpus.

METHODOLOGY

A three-tier approach was developed to facilitate thorough investigation of the speech corpora chosen for the study. The first tier used a schema devised by Ellen Douglas-Cowie of Queen's University, Belfast, and further refined for use in this study. It was used to assess the full range of speech features that might prove relevant. It contained elements relating to discourse structure, gesture, pitch, timing, voice quality, and articulation in addition to paralinguistic features. This type of open assessment addressed an area that has usually been ignored: that of auditory analysis of emotional speech. It also ensured that the study did not focus solely on a narrow range of acoustic measurements, as was common in the past, but maintained the flexibility to move into other areas of investigation if they proved relevant. The table below summarises the categories in the schema.

PITCH	Height
	Range
	Variability
	Pitch shape
TIMING	Overall rate
	Pausing
	Articulatory movement
	Rhythm
VOICE QUALITY	Overall muscular settings
	Laryn/pharyngeal settings
	Articulatory settings
VOLUME	Height
	Range
	Variability
	Patterns
PARALINGUISTIC	Reflexes
	Non-linguistic bursts
	Exclamations
	Other
ARTICULATION	Impairment
	Articulatory timing
	Articulatory settings

Figure 1. Table of schema contents

The second tier of investigation involved the creation of a digest of the findings of studies from the 1930s to the present. This digest extracted the speech measures that have proved most relevant, the emotions that we know most about, and also looked at the level of consensus on these issues. This provided a list of acoustic speech measures that formed the first stage of investigation, and ensured that the results could be evaluated in the light of themes that have emerged from prior research. Measures included: total time, mean intensity, mean pitch, minimum and maximum pitch, F0 range, formant frequencies from F1 to F4, syllables/second, % speech and % silence, and the frequency of stressed syllables. Pauses were broken down into two categories, silent and filled, and the average duration of these pause types was noted in addition to their frequency.

The third tier of analysis was the creation of transcriptions, both Interlinear Tonetic Transcripts, and transcripts indicating timing features shown on a time waveform. This ensured that both pitch patterning and timing and pausing issues were fully examined, and that the study had the potential to address the development of emotion over time, an issue largely ignored in many older studies of acted emotional speech.

Two databases were used to facilitate the study: The Belfast Structured Database [2], consisting of acted emotional speech, and the Belfast Naturalistic Database [3], consisting of naturally-occurring emotional speech taken from television interviews and interviews conducted by the creators of the database. Four emotional states, in addition to neutrality, were available for selection from the Structured Database: sadness, fear, anger and happiness. These were replicated in emotion selections from the Naturalistic database. A total of sixty-nine speech clips were selected for inclusion in the study. Nine clips were from the Naturalistic Database and consisted of two each of sadness, fear, anger, happiness and a single neutral clip. Sixty clips were chosen from the Structured Database. These consisted of groups of ten clips, consisting of two selections of sadness, fear, anger, happiness and neutrality from each of six actors. The actors were selected on the basis of genuineness and validity, as rated by naïve listeners for the purposes of another study. The actors rated most genuine and valid were used in the study.

Each speech clip from each database was assessed using the three-tier system. The results were then compared.

RESULTS

Results from the first tier of analysis imply that acted and non-acted emotional corpora differ quite dramatically in a number of areas.

On the paralinguistic level, non-acted speech is characterised by audible smile, which is three times more common than in acted speech. Acted speech is characterised by the presence of trembling lip and crying and sobbing, which are totally absent in the natural samples, is also evident.

In terms of pitch, natural speech is more likely to show evidence of both raised and lowered height, both narrow and wide range, changing variability and a predominant rising-falling pitch pattern. Acted speech is characterised by a relatively even distribution of pitch patterns between rising, falling and rising-falling patterns.

Timing is the area in which the differences between acted and non-acted emotional corpora are perhaps most evident in perceptual analysis. From a perceptual point of view, articulatory timing, or speech that is too fast of too slow, is much more likely to be thought incorrect in acted speech. Alternating articulatory timing is a feature of non-acted emotional clips.

The differences between acted and non-acted emotional corpora stood out starkly when the percentages for pauses were compared. There is greater variety in the types of pauses and pausing present in non-acted emotional speech. By comparison, acted emotional speech is bereft of pause variety: for example, disruptive pause is present in only 5% of acted samples, compared to a striking 22.2% of non-acted samples. Both over-long pausing and extremely short pausing are features of non-acted emotional speech. Rhythm also shows up clear differences between acted and non-acted material. 66.7% of non-acted samples were considered even, whereas only 16.7% of acted samples were.

The laryngeal settings of non-acted emotional speech are much more varied than acted emotional speech. Creaky voice is the most common, both in non-acted and in acted speech. At this stage, there is marked divergence. Breathy, harsh and whispery voice all are present in non-acted emotional speech. In acted speech, however, only whispery voice is present, and in very few samples.

Volume is not thought to be particularly relevant in the comparison of acted and non-acted emotional corpora. The only exception is the static variability that is a characteristic of non-acted emotional speech. It is worth noting that sporadic strong stress is present to almost the same degree in both acted and non-acted speech, but is only considered to be excessive in acted speech, in a mere 3.3% of samples.

On the perceptual level, articulation is also relevant. Mis-articulations are the only type of speech impairment present in both acted and non-acted speech, and the frequency is much greater in non-acted speech, at 33.3%. Stuttering and repetition are only evidenced in non-acted speech, and jerkiness in only acted speech. Excess sibilance is a feature of both types of emotional speech, but twice as common in non-acted as in acted speech.

These differences were also in evidence when individual speech clips from the acted and non-acted corpora were compared. The examples that follow are from a comparison of neutral speech from each corpus. Results show that some actors were capable of providing a sound clip that seemed remarkably close to the original non-acted clip. The best neutral acted speech clip, for example, scored similarly to the non-acted neutral clip in terms of overall clip length, syllables/second, percentage speech, percentage silence, average pause duration, and pause frequency.

	NON-ACTED	BEST ACTED
TOTAL TIME (SEC.)	28.370	27.688
SYLLABLES/SECOND	4.512	4.809
% SPEECH	80	80
% SILENCE	20	20
TOTAL PAUSES	10	9
AVE. PAUSE DURATION	0.588	0.627
TOTAL SILENT PAUSES	10	8
AVE. DURATION SILENT	0.588	0.545
TOTAL FILLED PAUSES	0	1
AVE. DURATION FILLED	0.000	1.291
PAUSE FREQUENCY	2.837	3.076
SILENT PAUSE FREQ.	2.837	0.461
FILLED PAUSE FREQ.	0.000	27.688

Figure 2. Non-acted and best acted comparison

The table compares a non-acted neutral clip with the best-acted neutral clip. This acted sample was one in a group of six. Whilst none of the other acted samples were as close to the non-acted clip, they did break down into three loose groups:

1. The acted sample closest to the non-acted clip – shown above.
2. Two acted samples with a similar rate of speech to the non-acted clip, but a much higher pause frequency.
3. The remaining acted samples, which had a much faster rate of speech and fewer pauses than the non-acted clip.

These disparities related largely to timing-related issues. The breakdown of the groups above indicates that, broadly speaking, acted and non-acted emotional corpora have more differences than similarities, and whilst individuals within the corpora may compare favourably, the fundamental conclusion remains that, as a whole, acted and non-acted emotional corpora do not compare.

Initial transcript analysis supports this conclusion. The best neutral acted clip was compared with the non-acted neutral clip, and indicates that whilst there may be broad similarities between the two clips in terms of instrumental timing measures, the actual breakdown of pauses is quite different. Although the number of pauses in each clip is very similar, the acted clip shows pausing at logically semantic moments in the text, where one might expect a full stop or a comma. The non-acted clip contains pauses that do not have a grammatical base, and which fall randomly throughout the text.

CONCLUSION

Although this research is still at an early stage, a number of conclusions about the relationship between acted and non-acted emotional corpora can be drawn. Comparative auditory, instrumental and transcript analysis of individual speech clips from two corpora, one of acted and one of non-acted emotional speech, show that there are few similarities. Comparison of individual emotional speech clips does show that the best acted material can appear similar on an instrumental level, but not on close contextual examination. This has two implications. It underlines the importance of using non-acted emotional speech corpora in applications, and it proves the usefulness of analysis above and beyond the instrumental level.

REFERENCES

1. Cowie, R., Douglas-Cowie, E., & Cox, C. Beyond emotion archetypes: Databases for emotion modeling using neural networks. *Neural Networks 18* (2005), 371-388.
2. Douglas-Cowie et al. (2000). Belfast Structured Database.
3. Douglas-Cowie et al. (2000). Belfast Naturalistic Database.

Evaluation of expressive speech resynthesis

Nicolas Audibert[1], Damien Vincent[2], Véronique Aubergé[1], Olivier Rosec[2]

[1] Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
{audibert, auberge}@icp.inpg.fr

[2] France Telecom, R&D Division
Lannion, France
{damien.vincent,olivier.rosec}@francetelecom.com

ABSTRACT

Expressive speech is intrinsically multi-dimensional. Each acoustic dimension has specific weights depending on the nature of the expressed affects. The quantity of expressive information carried by each dimension separately (using Praat algorithms) has been perceptively measured for a set of natural mono-syllabic utterances (Audibert et al, 2005). It has been shown that no parameter alone is able to carry the whole emotion information. These selected stimuli, expressing anxiety, disappointment, disgust, disquiet, joy, resignation and sadness were resynthesized with an LF-ARX algorithm, and evaluated in the same perceptive protocol extended to the three voice quality parameters (source, filter and residue). The comparison of results between natural, TD-PSOLA resynthesized and LF-ARX resynthesized stimuli (1) globally confirms the relative weights of each dimension (2) diagnoses local minor artifacts of resynthesis (3) validates the efficiency of the LF-ARX algorithm (4) measures the relative importance of each of the three LF-ARX parameters

Keywords

Expressive speech, affects, resynthesis, prosodic dimensions, LF-ARX model, evaluation.

INTRODUCTION

In order to study the expressive speech, whether for theoretical purposes or for applications in synthesis or recognition, one first has to face the fundamental problem of the acoustic dimensions of the affective prosody. From the old and unsolved debate about the specificities of some acoustic dimensions possibly devoted to particular affects, it can be at least retained that all prosodic dimensions, i.e. F0, intensity, “voice quality” and duration, must be tracked to model all kinds of affects (moods, emotions, attitudes, …). A central open question is in particular to understand if the voice quality must be considered globally as a single dimension, or should be described as several dimensions. The glottal articulatory-to-acoustic modes (breathy, creaky voice, etc.) are quite well described [9], but even if some studies link one of these modes to some affects characteristics (e.g. breathy to intimacy [9] or care [6]), most studies globally describe the voice quality globally by a great number of acoustic parameters. In spite of its complexity, expressive speech inversion using the LF model [7] theoretically makes possible the characterization of the signal without using redundant and not significant acoustic parameters. Our aim is to evaluate how the LF model can encode affective information in real spontaneous data, not globally, but specifically for each dimension (F0, intensity, duration, voice quality), by comparing for each dimension the LF coded stimulus and the original stimulus.

The methodology used is based on copy synthesis, i.e. an acoustic analysis of reference stimuli used as an input for the synthesis of new stimuli, followed by a perceptive evaluation. Analyzed parameters may be voluntarily altered, and either the whole set of parameters or only part of them may be used. To study expressive speech, Gobl et al. [8] evaluated the role played by voice quality in emotional expressions using stimuli synthesized from a reference glottal flow waveform modified to express different voice qualities. In another experiment [5], F0 variations and durations analyzed from expressions of anger, happiness and sadness as well as neutral expressions were applied to diphones carrying the same emotional expressions in concatenative synthesizers. The authors concluded from the perceptive evaluation of stimuli built with prosodic parameters not matched to diphones that anger was mainly carried by diphones and sadness by prosodic parameters, when no clear pattern could be found for happiness. In other studies (e.g. [3]), stimuli were synthesized from multiparametric measurements to evaluate the relevance of the extracted parameters for the perception of emotional expressions.

In a previous study [2], monosyllabic stimuli carrying 8 emotional expressions were used as a basis to generate synthetic stimuli using the Praat software [3], by projecting analyzed parameters separately. The perceptive evaluation of the generated stimuli revealed that F0 contours carried more information on positive expressions, voice quality and duration carried more information on negative expressions, and intensity brought no information when used alone. However the synthesis method did not make possible the separate evaluation of the influence of voice quality vs. duration. The aim of the present study is to replicate a similar evaluation on stimuli generated from the same set of natural stimuli with an LF ARX algorithm [12] able to process voice quality and duration independently, and to compare perceptive results with those previously obtained.

EXPERIMENTAL FRAMEWORK FOR COPY

SYNTHESIS

Speech model

Many speech production models hypothesize that a speech signal can be considered as the result of passing an excitation through a linear filter. In this source-filter approach, the source part refers to the so called glottal flow derivative (GFD) which corresponds to the signal produced at the glottis and integrating the effect of lip radiation approximated by a derivation, while the filter models the vocal tract resonances. When voiced sounds are produced, the vocal fold vibration results in a quasi periodic GFD for which classical models exist. The model used in this paper is the LF model [7], which allows a parameterization of the shape of GFD waveform with 3 parameters. A stochastic component is also present to model noise-like effects (irregularity of the GFD, fricative noise, etc...). Given the above assumptions, a voiced sound *s(n)* can be modeled by an ARX (Auto Regressive eXogenous) process defined as:

$$s(n) = -\sum_{k=1}^{p} a_k(n).s(n-k) + u(n) + e(n)$$

where *u(n)* and *e(n)* respectively denote the deterministic (LF) and the stochastic parts of the GFD, and where $a_k(n)$ are the coefficients of the filter of order p characterizing the vocal tract.

Given this speech model, the analysis falls down to estimating the vocal tract filter, the F0, the energy coefficient, the LF waveform parameters and a residual component. The joint estimation of this information is not straightforward as the optimization over the LF parameters is not a linear problem. However when the LF parameters are known, the estimation of the filter and the residue can be achieved by least square methods: an efficient method was thus proposed for solving this estimation problem by an exhaustive search over a space of quantized LF waveforms followed by local optimizations 12.

Implementation issues for copy synthesis

Let us consider a text message uttered by a speaker in two configurations: one considered as neutral and the other one carrying an emotional content. As we are interested in identifying the relevant correlates for conveying emotion, copy synthesis will essentially consist here in replacing some of the parameters in the neutral utterance referred to as the source stimulus by their counterparts in the emotional target stimulus. For this purpose, two tasks are necessary: first an alignment procedure so as to map events between source and target stimuli and second a synthesis algorithm which enables the transformation of the desired correlates.

As the alignment procedure is phonetically constrained, a prerequisite is that both stimuli have the same phonetic content, the phoneme segmentation being available. Then, the matching between source and target frames is done by matching the phoneme boundaries and by relating the analysis instants within each phoneme by means of a linear interpolation. It is worth noting that problems can occur when the voicing information of source and target stimuli are different. However, after careful inspection of all speech signals, we did not find this kind of mismatch. During the synthesis step, once the alignment is done, the synthesis instants can be determined by an algorithm similar to the one used for TDPSOLA based prosodic modifications 9. Thus, this algorithm provides for each synthesis instant a pair of analysis frames respectively from the source and target stimuli, making the copy synthesis straightforward.

Generation of synthetic stimuli

The 10 stimuli used as a basis for the copy synthesis were the monosyllabic stimuli carrying emotional expressions as used in a previous resynthesis study carried out at ICP 2, extracted from the E-Wiz / Sound Teacher corpus [1] and perceptively validated [11], phonemes boundaries being manually labeled. Satisfaction was discarded from this set as the aforementioned framework failed to give a sufficient good quality for generated stimuli, this problem being under investigation. Thus, this set was restrained to 7 stimuli expressing anxiety, disappointment, disgust, disquiet, joy, resignation and sadness as well as neutral expressions on French monosyllabic words: [ʁuʒ] and [sabl].

From the analysis of different stimuli, the 6 following sets of parameters could be set independently to the value either of the expressive stimulus or of the corresponding neutral expression: F0, intensity, phonemic duration, source, residue and vocal tract filter. All 64 combinations of the 2 possible values of these 6 sets of parameters were systematically generated from each of the 7 expressions. However, only 7 synthesis conditions were selected for the perceptive evaluation, labeled according to the parameters of the target stimulus used, other parameters being extracted from the neutral expressions: (i) *'control'*, obtained by applying all the parameters extracted from the expressive stimulus (ii) *'VQ & duration'* (iii) *'VQ'* (iv) *'source & residue'* (v) *'source'* (vi) *'duration'* and (vii) *'F0 & intensity'*. The selected subset thus counts 49 stimuli. 2 additional stimuli were selected in the control condition, generated as a copy synthesis of the neutral expression stimuli, for a total of 51 stimuli. The control, *'VQ & duration'* and *'F0 & intensity'* conditions were selected to enable direct comparison with previous results 2.

PERCEPTIVE EVALUATION AND RESULTS

The 51 generated stimuli were perceptively evaluated by 25 judges (7 male, 25 female, aged 25.7 in average) at ICP, in a soundproof room with high quality headphones, with 3 presentations of each stimulus. Stimuli were presented to each judge in a different random order, the same stimulus being not presented twice consecutively. Stimuli presentation and judges' answers recording were performed using an automated interface: judges had to select either an expression within the 7 proposed (anxiety, disappointment, disgust, disquiet, joy, resignation and sadness) or the neutral expression. Judges were also asked to rate the perceived emotional intensity on a 1-10 scale. The high Cronbach's alpha value (α=.92) indicates that answers given by different judges are coherent with each other.

Results were then distributed into confusion matrices and analyzed separately for each synthesis condition. As rated emotional intensities, significantly correlated to identification scores (r^2=.889), do not bring additional information, the analysis presented hereunder focuses on identification scores. However consequences of the inter-judge effect on the expression of disgust could be observed as in [2], since it was attributed the highest emotional intensities both in control and *'duration'* conditions though other expressions were better identified.

As the confusions were similar to those observed in [2] in control condition, the same clustering was applied: anxiety and disquiet were grouped together, as well as resignation, disappointment and sadness, while joy, disgust and neutral remained separate categories. As most of the relevant information appears on the matrices diagonals (i.e. identification scores) after clustering, data were converted to right or false answers and normalized to make possible further statistical evaluation. Transformed data were tested using an ANOVA of repeated measures with synthesis condition and expression as fixed factors, showing a significant effect (p=.01) of condition, expression, and of the condition*expression interaction. ANOVAs of repeated measures were computed for each synthesis condition, revealing a significant effect (p=.01) of the expression for all conditions except control. Inter-expression contrasts significance (p=.01 when not specifically stated) was also systematically tested for each condition, as well as between different conditions for a given expression.

Table 1 summarizes identification scores for each label and each synthesis condition after clustering in the present evaluation (labeled 'ARX'), together with scores obtained for stimuli synthesized with Praat [2] (labeled 'Praat') when comparison is possible. As joy and satisfaction were confused in this study, confusions of joy with satisfaction were included in the calculation of the identification score of joy. Identification scores of natural stimuli for each cluster of expressions, derived from [11], are also presented in this table (labeled 'natural'). Identification of natural stimuli appears to be higher than identification of synthetic stimuli, except for the expression of disgust. However this difference could be explained as a consequence of the inter-judge effect observed in 11. As the structure of collected data does not make possible a statistical evaluation of differences between results obtained in these 3 studies, comparisons across sets of results are only qualitative.

		Natural	control	F0+int	VQ+dur.	duration	VQ	source+res	Source
Joy	Praat	80.9%	70.8%	42.5%	6.7%				
	ARX		77.3%	58.7%	30.7%	0%	10.7%	4%	4%
sadness, resign. disapp.	Praat	82%	59.6%	26.7%	55.8%				
	ARX		56%	27.1%	52.9%	56.9%	44.4%	44%	41.8%
Anxiety disquiet	Praat	76.1%	55.6%	21.4%	47.2%				
	ARX		67.3%	40.7%	60%	46%	46%	35.3%	24%
Disgust	Praat	55.7%	61.7%	3.3%	34.2%				
	ARX		70.7%	1.3%	42.7%	49.3%	8%	5.3%	1.3%
Neutral	Praat	66.1%	31.7%						
	ARX		52.7%						

Table 1: Identification scores compared with those obtained in 2 and 11, both after clustering.

Considering identification scores of the present study, a first observation is that major trends observed in 2 are confirmed. Indeed manipulated stimuli were generally not identified as well as control stimuli. However a few exceptions appear: the identification score of the sadness, resignation and disappointment expressions, as well as of the anxiety and disquiet expressions, were not significantly different in *'VQ & duration'* vs. control condition. In control condition the scores show a pattern similar to 2, most of the expressions being slightly better identified, with a large improvement for the neutral expression (52.7% vs. 31.7%). In *'F0 & intensity'* condition most of the expressive information on joy is retained, though the identification score is significantly lower than in control condition (58.7% vs. 77.3%), since joy is significantly better identified than other expressions. Most of the affective information on anxiety and disquiet, as on sadness, resignation and disappointment, was retained in the *'VQ & duration'* condition. However this tendency is stronger for anxiety and disquiet (60% vs. 47.2% in [2]), suggesting that voice quality of these expressions was better retained with LF-ARX vs. Praat. In this synthesis condition the part of the retained information on disgust (identified at 42.7% vs. 70.7% in control condition) is comparable to previous observations. It was already shown with intensity that a 'weak' parameter may significantly improve the identification of an expression when combined with other parameters 2. However this quite low identification, whereas F0 and intensity carry very little information, seems to confirm the observation that disgust is more sensitive to manipulations than other expressions 2.

On the other hand the relative influence of voice quality vs. duration was evaluated. In *'duration'* condition, sadness, resignation and disappointment were as well identified as in control condition (no significant difference), though these expressions were identified significantly fewer but much over chance level in *'VQ'* condition. For the expressions of anxiety and disquiet, duration and VQ appear to carry the same amount of affective information (both identified at 46%). Eventually, phonemic duration carries most of the information on the expression of disgust (no significant difference when compared to *'VQ & duration'* condition).

Moreover the relative weights of the parameters used for the modeling of voice quality by the LF-ARX algorithm were evaluated by comparing scores obtained in *'VQ'*, *'source & residue'* and *'source'* conditions. On sadness, resignation and disappointment, as well as joy, the source parameters appear to carry the whole information on voice quality: these expressions were not significantly better identified when residue and filter information were used vs. source only. For the expression of disgust this difference is hardly significant (p=.05), the score in *'VQ'* condition remaining below chance. On the other hand anxiety and disquiet were significantly better recognized in *'source & residue'* vs. *'source'* condition, and in *'VQ'* vs. *'source & residue'* condition.

DISCUSSION

The score obtained for the expression of joy with LF-ARX in *'VQ & duration'* condition (30.7%) should a priori be directly compared to the score in *'VQ'* condition (10.7%), but 2 surprising results appear: this expression was not identified at all (0%) in *'duration'* condition, though the difference between *'VQ'* (10.7%) and *'VQ & duration'* (30.7%) is far above zero. On the other hand this expression in *'VQ & duration'* condition reached only a score of 6.7% using Praat *vs.* 30.7% using LF-ARX. This led us to look for a possible artifact, so we noticed that an unexpected very short, mean energy "closure" noise appears at the beginning of the stimulus generated with LF-ARX in *'VQ & duration'* condition. As the copy synthesis algorithm was originally designed for "clean" speech, where segments labeled as silence actually correspond to a silence, silence segments are generated by copying the silence part of either the source or target stimulus. Since this noise appears in the neutral expression, it was automatically copied to the generated stimulus. We assume that this noise was interpreted by judges as a laughter cue, making the identification score higher. This unexpected artifact points out a very interesting phenomenon, since it shows that the characterization of affects in speech cannot be reduced to quantifying (by qualifying) the signal information.

A more relevant comparison would thus be between *'VQ'* condition with LF-ARX (10.7%) and *'VQ & duration'* condition with Praat (6.7%), as we want to compare those 2 algorithms. Another artifact might have lowered the score with Praat (6.7%): to generate this stimulus, the relative intensity contour of the neutral expression was applied and the signal was scaled to reach the target stimulus overall energy level. As local intensity values are not controlled, the generated expression shows lower intensity at the end than the neutral expression, though their mean intensities are equal. We assume that this final low intensity was interpreted by judges as incongruent with a joyful expression. It can thus be expected that, when replicating this experiment with "clean" stimuli, expressions of joy generated with LF-ARX in 'VQ' and 'VQ & duration' conditions would get the same score, around 10%.

ACKNOWLEDGMENT

This work was conducted jointly by ICP Grenoble and France Telecom R&D, partly within the NoE Humaine.

REFERENCES

1. Aubergé, V., Audibert, N. and Rilliard, A., 2004. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, 179-182.
2. Audibert, N., Aubergé, V., and Rilliard, A., 2005. The prosodic dimensions of emotion in speech: the relative weights of parameters. *Interspeech 2005*, 525-528.
3. Bänziger, T., Morel, M., and Scherer, K. R., 2003. Is there an emotion signature in intonational patterns? And can it be used in synthesis? *Eurospeech 2003*, 1641-1644.
4. Boersma, P., and Weenink, D. Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat
5. Bulut, M., Narayanan, S. S. and Syrdal, A. K., 2002. Expressive speech synthesis using a concatenative synthesizer. *7th ICSLP*.
6. Campbell, N., and Mokhtari, P., 2003. Voice Quality: the 4th Prosodic Dimension. *15th ICPhS*, 2417-2420.
7. Fant, G., Liljencrants, J., and Lin, Q., 1985. A four-parameter model of glottal flow. *STL-QPSR* (4), 1–13.
8. Gobl, C., and Ní Chasaide, A., 2003. The role of the voice quality in communicating emotions, mood and attitude" *Speech Communication* (40), 189–212.
9. Laver, J., 1980. *The phonetic description of voice quality*. Cambridge: Cambridge University Press.
10. Moulines, E. and Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modifications of speech. *Speech Communication* (16), 175–205.
11. Rilliard, A., Aubergé, V. and Audibert, N., 2004. Evaluating an Authentic Audio-Visual Expressive Speech Corpus. *4th LREC*, 175-178.
12. Vincent, D., Rosec, O., and Chonavel, T., 2005. Estimation of LF glottal source parameters based on arx model. *Interspeech 2005*, 333-336.

Body and music. An annotation scheme of the pianist's multimodal behaviour

Isabella Poggi

Dipartimento di Scienze dell'Educazione
Via del Castro Pretorio 20 – 00185 Roma – Italy
poggi@uniroma3.it

ABSTRACT

The paper presents a research on the multimodal behaviour of a pianist during performance, and proposes an annotation scheme for this behaviour. The research data, aimed at a first testing of a theoretical model of the pianist's mind, shed some light on the relation of trunk, head and face movements with hand movements during piano performance, and on their expressive import as to the pianist's cognitive, emotional and motor processes. The annotation scheme is presented and some uses of it are proposed for research on music performance.

Keywords

Annotation scheme, Music performance, pianist.

INTRODUCTION

Interaction linked to music performance has been analysed with methods of conversational analysis [2], [3], [6], while the gestures of classic orchestra conductors were analysed in terms of their visual metaphors [1].

But analysis of multimodal communication in music performance can also be carried on through more analytical annotation systems. An example is the classical orchestra conductor's face, that was analysed [4] through a variant of the "score of multimodal communication" [5]. In this work I present an annotation scheme for the analysis of the pianist's multimodal body behaviour.

A TOP-DOWN AND A BOTTOM-UP APPROACH

Research implies testing theoretical hypotheses by collecting and analysing data. But the very elaboration of a hypothesis often comes about thanks to induction starting from data observation. So any investigation uses both a top-down and a bottom-up approach, one starting from theory and one from data. In a research aimed to test a theoretical model of the cognitive and emotional processes holding in a pianist's mind while playing the piano, I had the chance to tune up a variant of the "score of multimodal communication" [5], that can be used to transcribe and analyse multimodal behaviour during music performance. In the following I present the theoretical model proposed, the research aimed at its testing, and finally the annotation scheme of the pianist's body tuned up thanks to this study.

THE PIANIST'S BODY AND THE PIANIST'S MIND

While playing the piano, three kinds of processes take place in the pianist's mind.

- Cognitive processes. While playing, cognitive processes of attention, perception, and memory are at work.

Attentional processes include, for example, the pianist being in concentration on the music to play, that is, his focusing of attention aimed at performing at his best. On the contrary, the pianist can sometimes be quite relaxed, for example when the piece to play is well mastered: and in this case he can suspend his tension and concentration. Attention is also implied, for example, in "caution", when the pianist plays very softly, like in touching something fragile, something not to be spoiled with too vigorous a touch.

Perception is obviously at work while the pianist is playing. Of course, perception relevant to his playing is especially acoustic perception, both in hearing other players' music, to start playing, to be consonant with others' music, and in hearing his own music, to have a feed back of how he is playing.

Memory: while playing, the pianist has auditory images of the sounds to play, and visual images not only of the score to play, but also of objects or landscapes: any simply visual or synesthaetic image linking sound to vision. Moreover, tactile and spatial memory of where tunes are or how they are felt by hands. Finally, very much linked to these, the procedural memory of hands and feet movement is central in the pianist's cognitive processes, since it governs his motor processes.

- Emotional processes. In a pianist we can distinguish *felt emotions* and *enacted emotions*.

The pianist may feel real emotions while playing, but he also must recite emotions, feel as if he were feeling some. A pianist, and not only a professional one, is like an actor on the stage, who must induce emotions in himself: the emotions to be impressed in music, and thus to be transmitted to listeners. At the same time, a professional pianist generally has a capacity to manage his emotions, just as an actor does, in such a way as not to be overwhelmed by them while playing.

Two types of *felt emotions* can be distinguished:

- *process emotions*, the emotions felt during and about the very process of playing: the positive emotions of relaxation and flow (a sensation of being doing something beautifully and easily), and the negative ones of tension, or fear of making mistakes in performance.
- *outcome emotions*, the emotions felt about the outcome of one's playing: disgust or shame for making a mistake, pleasure, satisfaction or ecstasy for how sweet it sounds.

Enacted emotions can be:

- *meaning oriented* emotions, the ones the pianist must simulate or induce in himself in order to exhibit them through music: for instance, feeling sad to play a sad music.
- *movement oriented* : the emotions the pianist needs to simulate because their expression is linked to a particular movement or manner of movement. Typically, for example, a pianist frowns when he has to play very loud notes: frowning is an expression of anger, and anger makes you so strong and energetic as to touch keys with more strength. Here the emotion, and its consequent expression, is functional to the quality of sound.

- Motor processes, relative to the movements to perform and to the manner in which they should be performed.

Music is produced by movement; among the parameters of music, melody, rhythm and harmony could be seen as produced by hand movements (which tunes are touched by fingers and when), while parameters like timbre, tempo, expression, intensity as determined not by the movements themselves, but by the way in which they are performed – their "manner of movement".

In a pianist, the combination of cognitive and emotional processes determines the way in which motor processes work in producing music.

AN ANNOTATION SCHEME OF THE PIANIST'S BODY

This is my hypothetical model of the pianist's mind. But how can we know if it is an adequate model? An empirical evidence would be to find out external signals of these processes. So, leaning on the principles of the "score of multimodal communication" [5], I tuned up a specific annotation scheme for the analysis of the pianist's multimodal behaviour. The pianist Marcella Crudeli, while playing the piano concert in A major K 488 by W.A.Mozart, was videotaped during a rehearsal and a public concert, and her multimodal behaviour was analysed through this particular variant of the "score" (Table 1). Later the results of the analysis were shown to the pianist, who provided some further comments and generally agreed about the interpretation of data provided.

Table 1

Bar	***1*** *100*	***2*** ***Goal / Meaning***	***3*** ***Goal Type / MeaningType***
1. *notes*	B E D# C B A# A G		
2. *time*	1.20		
3. *trunk*	Shoulders raised and backward	I retract, I do not want → Sorrow	CP → Mea E
4. *head*	Shakes head slowly	No no = I reject → sorrow	CP→ Mea E
5. *eyebrows*	Frowns, then raises inner eyebrows	I am worried + I am sad	Mea E + Mea E
6. *eyes*	Squeezes eyes, then opens eyes	Help play sharp A (# A)	HM
7. *mouth*			

In this study the analysis is carried out only on the pianist's body, not on her hands and feet, that obviously do the technical job of playing.
The analysis is necessarily carried on by making reference to the score of the music performed and by examining fragments bar by bar (sometimes even note by note). In the first column you write the bar number, and in the lines 1 through 7, respectively, the following information:

1. the notes of that bar, either in musical notation or with their names;
2. the time at which the passage occurs;
3. – 7. the modalities taken into account for the analysis; in this study, respectively, trunk, head, eyebrows, eyes and mouth.

For each modality, the analysis is written down in the first three columns. In col 1. you write a description of the movements performed by that part of the body; in col. 2, an interpretation of that movement: if it is judged a non-communicative action, you simply write its goal, while if it seems to have expressive or communicative goals, you write down the information it aims to provide, that is, its meaning. An action is defined as "communicative" if it has a conscious, unconscious or tacit goal, or a goal determined by social or biological function, of having some other Agent assume some belief [5]. An "expressive action" is a type of communicative action that concerns information about the Sender's mental states (not about the external world), and whose Sender is not aware either of his goal of communicating or of the signal produced or of both. An example of "communicative action" is when the Pianist, during the rehearsal, nods to the orchestra to praise for how it is playing. If she frowns in concentrating before starting, this is an "expressive action". If she

moves her head up and down rhythmically to accompany the rhythm of the music she is playing, this is a "non-communicative action".
In col. 3 you classify the goal or meaning of col.2 in terms of a typology of the movements performed (see Table 2 below).
In col. 2 and 3 you can provide an analysis not only of the literal meaning or of the apparent goal of each movement, but also of its indirect meaning, or its superordinate goal.
For example, this is the analysis of head behaviour in Table 1, row 4. At bar 100, the pianist shakes her head slowly (col. 1), like if saying "no no", a performative of rejection; but what she rejects, as confirmed by the context, namely the meaning "I am sad" (line 5, col.2) provided by the raised inner eyebrows (line 5, col.1), lets you infer that she feels sorrow; sorrow is well something one rejects. So, in line 4, col.2, the literal meaning is "I reject", but the indirect meaning is "I feel sorrow". The analysis of these two meanings in line 4., col. 3, classifies, respectively, the literal meaning as CP (a Communicative act, namely a Performative), and the indirect meaning as a MeaE (Meaning oriented emotion): the rejection the pianist is communicating is a way to express her sorrow, which thus se induces in the sounds she is playing.
On the right of the first three columns, new bars can be written down and analysed in the same way. Moreover, one can add other lines underneath to provide a parallel analysis of the multimodal behaviour of another pianist at the same bar, or of the same pianist in rehearsal vs. concert. This allows to compare the different behaviours and to state if the same body behaviour systematically co-occurs with the same hand movements (the same notes to play) across performances and across pianists.

BODY MOVEMENTS WHILE PLAYING THE PIANO

This first research allowed to state a typology of the possible functions of the pianist's body movements, to be used in the classification of col. 3 (Table 2). The types of actions or meanings provided by the pianist's trunk, head and face are the following:

- Communicative acts. Sometimes the pianist, even during performance, communicates something to specific people, and since music prevents verbal communication, this occurs through gaze, head movements or facial expression. In some cases only the performative of a communicative act is conveyed (CP), and the propositional content is to be understood from context: for example Marcella Crudeli, with a head nod, eyebrows raised and a smile, praises the orchestra. But sometimes also the propositional content of the communicative act is explicitly conveyed by trunk or face (CPc): with fast movements she incites the orchestra to go faster.

Table 2

Communicate	Performative	CP
	Performative + Content	CPc
Express or communicate Emotional Processes	E/C Process Emotions	PE
	E/C Outcome Emotions	OE
	E/C Meaning Emotions	MeaE
	E Movement Emotions	MoE
Express Cognitive Processes	Express Attention	EA
	Express Perception	EP
	Express Memory	EM
Help Motor Action	Help Melody	HM
	Help Rhythm	HR
	Help Harmony	HH
Help Manner of Movement	Help Tempo	The
	Help Timbre	Hti
	Help Intensity	HI

- Communication and expression of emotional processes. Given our definition of "expression", information about the music can only be communicated, while information about the pianist's emotions or cognitive processes can be either communicated (if the pianist deliberately and consciously signals it) or expressed (if it leaks instinctively or without awareness). Process emotions (PE) and outcome emotions (OE), since they are felt, can be either expressed or communicated. Among enacted emotions, those meaning oriented (MeaE) will presumably be communicated, if they are consciously simulated, while movement oriented emotions (MoE) may be simply expressed: the pianist might be conscious of the goal of making that movement, not of simulating the emotion *in order to* make that movement. An example of Process emotion in Crudeli's data is her shaking her head rhythmically and smiling, showing mirth about her own playing; she displays an Outcome emotion of pride and satisfaction about how she played by raising external eyebrows and smiling. By raising inner eyebrows she expresses a Meaning oriented enacted emotion of sorrow, and she frowns enacting anger – a Movement oriented emotion – when playing loud notes.
- Expression of cognitive processes. The pianist can express her mental states of attention (EA), perception (EP), and memory (EM). Lowered eyebrows show attention and concentration; head down with ear close to the tunes shows the pianist is listening to the sounds played; a signal of memory processes (presumably typical of novice pianists, not of professionals in concert, might be eyes up while trying to remember.
- Accompaniment of hand motor actions. In some cases head, face and trunk movements do not have a communicative or expressive import: they are "simply" movements. But since they are often synchronous and analogous to the movements of the pianist's hands, they even seem to "help" them to move better. An objection to this claim is that, were this so, pianists that move their body more would be better pianists. Which is not. Actually, different schools of piano technique exist, some encouraging body expression, others composure; but also people of different personalities may be more or less keen to body motion, and yet be equally good

pianists. However, saying that one can be a good pianist even without so much moving his body does not exclude that moving head or trunk at the same time as hands does help. After all, if work songs helped Negroes to work better, why should body movements not help hand movements? Within the movements that directly help the hand motor actions, we can distinguish those that help the actions aimed, respectively, at producing melody, rhythm, and harmony. Movements that help melody (HM) exploit a spatial analogy: direction and intervals of trunk head and face movements are analogous to those of hands. For example in Crudeli's data, at bar 209, the pianist's right hand makes a circular movement to play the tunes E, #G #F M; at the same time, her head makes an analogous circular movement. Or again, in playing an acciaccatura from an E to another E an octave higher, as she moves her right little finger up, at the same time she moves her external right eyebrow up! In movements that help rhythm (HR), their rhythmical structure is analogous to that of hand movements: at bars 210 – 211, head shakes rhythmically every two quartines. In movements helping harmony (HH) an analogy holds with the visual image of harmony relations: the pianist, for example, plays a piece around the tonic and then, while playing it again a third lower, lowers her eyebrows. Within the movements helping the manner of movement, some help tempo, others timbre, others intensity. Movements helping tempo (HTe) use the same tempo of the hands; like in the fast head shaking that accompanies fast hand movements; to help timbre (HTi), the tension of body movements is analogous to that of hands: like in raising the external parts of eyebrows which helps a tense timbre; to help intensity (HI), the movement or position of trunk, head or face allows or favors the intensity of hand movements; like in raising whole eyebrows, that helps making a delicate sound. Actually, in some cases it is not easy to tell if some action should be classified as Help Manner of Movement or as Movement Oriented Action, since as we said the manner of movement is very much linked to emotion expression [7].

CONCLUSION

The annotation scheme presented here can be used in research on music performers. After analysis through it, the different distribution of the functions of body movements illustrated above could be computed in different pianists (amateur and professional, with different personalities or coming from different technical schools, and playing different types of music); this could allow a better understanding of the role played by a musician's body in making music and expressing its meanings. In any case, it could once more show us the subtle synchrony and harmony of our body.

ACKNOWLEDGMENTS

LREC participation supported by HUMAINE (European Project IST- 507422). I am indebted to M[a]. Marcella Crudeli and to M[o]. Fabrizio Santi and the orchestra "Res Musica" for allowing me to videorecord and analyse their performance.

REFERENCES

1.Boyes Braem, P. and Braem, T. Expressive gestures used by classical orchestra conductors. In C.Mueller and R.Posner (Eds.), Proceedings of the *Symposium on The Semantics and Pragmatics of everyday Gestures*. Berlin: Weidler, 2004, 127-143.

2.Duranti, A.and Burrell, K. Jazz improvisation: a search for hidden harmony and a unique self. Ricerche di Psicologia 3, 27, 71-101.

3.Haviland, J. Batons in the hands, and batons in the air: Creating a gestural environment. 9th International Pragmatics Conference, Riva del Garda, 10-15 July 2005.

4.Poggi, I.: The lexicon of the conductor's face. In P.McKevitt, S. O' Nuallàin, Conn Mulvihill (Eds.) "Language,Vision, and Music. Selected papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, 1999". Amsterdam: John Benjamins, 2002, 271-284.

5.Poggi, I. and Magno Caldognetto, E.: *Mani che parlano. Gesti e psicologia della comunicazione*. Padova: Unipress, 1997.

6.Streeck, J and Oshima, S. Hip-hop gestures. The conduct of M.C. Paper at the Interacting Bodies Conference, Lyon, 15-18 June, 2005.

7.Wallbott., H.G. Bodily expression of emotion. European Journal of Social Psychology, 1998, 28, 879–896.

Auto-annotation: an alternative method to label expressive corpora

Véronique Aubergé,

Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
Veronique.Auberge @icp.inpg.fr

Nicolas Audibert

Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
Nicolas.Audibert@icp.inpg.fr

Albert Rilliard

Institut de la Communication Parlée
CNRS UMR 5009, Grenoble France
Albert.Rilliard@icp.inpg.fr

ABSTRACT

We discuss about the possible complementary quality of two approaches of labelling of expressive corpora: the main one which is the annotation by human experts, and an alternative one, presented here, which is based on auto-annotation by the subjects him-self.

Keywords

Emotion annotations, expressive corpus.

INTRODUCTION

One major problem is still to label the authentic expressive corpora. This problem is first to agree with the kinds of descriptors used to label affects in the corpora. Some hypothethico-inductive methodologies, like proposed by Cowie et al (2004) solve this problem by strong theoretical hypotheses on the nature of emotions, from which could be deduced labels, and labeling paradigm. Another question which can be, or not related to the first one, is to put the labels on recorded corpora by showing/listening these audio, visual or audio-visual corpora. The mainly used approach is to label corpora by human experts. We discuss here about the possible limits of this approach and about a complementary paradigm that we applied on a large multi-speakers corpus, and which is based on the auto-annotation labeling of emotions by the speaker him-self.

EXPERTS LABELING

Some elaborated methods have been held (Campbell, 2004), in particular in verifying, statistically, a posteriori, the coherence of annotation by different experts (see for example Devillers et al, (2005)). In the case of real-life and complex emotional data, it is surely the more relevant methods, for which the efficiency has been shown.

In such a method, a possible artifact if for the "expert" to take part of the situation, since to observe an interaction between human is by itself an ecological situation: in such hypothesis, this expert has competence to label, not because he is in a meta-task, outside the situation, but because he uses his own human emotional "normal" competences, built also by his own previous experiences and personality. Qualifying a human of expert, in using this human, not like a human, but like a "meta-observer" is classical in human productions annotation, mainly for language annotation, that is phonetics or linguistics. It is justified for emotions annotation when this annotation is based on a theory, which gives the "rules" and the tools to the human to be over his human abilities. But even when the human is really trained to be an expert, it is well known that even for phonological annotations, which is by definition a consensual code, the transcription of corpus by experts are very variants and subjective to the own phonological experience of the speaker (see SAM EU projects reports).

AUTO-ANNOTATED/NAIVE LABELING

We have experimented an alternative method, that is not as generic as the "by experts" methods, but that is devoted to be very close to the recorded subject in order to be as most as possible precise in the timing of labeling, without to be connected to any theoretical hypotheses on labels. The main advantage of auto-annotating methods is to avoid the risk of the expert to "take part" of the situation.

Considering that our need is to label very fine and timed parts of the corpus, we tried to use an annotation method where the subjectivity of labeling is the subjectivity of the subject him-self.

It has been shown early (Wallbott et Scherer, 1989) that the emotional experience of a subject is powerful and could be partially explained by new hypotheses of emotion theory. We experimented an auto-annotated method on several conditions (1) to imagine a scenario with which to trick many subjects in the same way in predicting roughly the subjects behaviors, in order to build further labels classes. The variation of emotional reactions of the subjects must depend only on the psychological profile of the subject – in E-Wiz/Sound Teacher, the scenario is an human-machine interaction, the trick is held by a wizard of Oz (Aubergé et, 2004); (2) to induce strong and simple variations of emotional states in order for the subject to find an easy way to label the combination and the degree of felt emotions (3) to request each subject just after the trick in a systematic way (4) to record very good quality of gestures, faces and speech to give to the subject for self-expertise, associated with bio-physiological signals for an a posteriori verification of labeling.

Since mainly studies have shown the accuracy of the auto-biographical memory of emotions, the main risk of this method is on one hand the limit of the sincerity of the subjects, and on the other hand the reliability between these labels and reference labels.

THE NATURE OF SELF-LABELS

Each session of Sound Teacher is around 40 minutes long. The subject can see the video tape (on which the precise time is written) of his session as many times as needed. He a specific form timed on his performance, on which are written in detail the different steps of the scenario, with free places on which the subject can label following his choice : they were told what they felt (not only what they expressed) to label with draws, signs, smileys, words, stories... All the 17 subjects said that the self-labeling was very easy. All of them have seen the video tape only one time. They did not mention having any problem to choose the kind of labels to use. Part of the subjects are actors, from a like Actor Studio method (which main feature is to remember a felt experience), and they however did not report to have been influenced by their actor training to remember and choose labels to their Sound Teacher experience.

Of sure, such a method has some problems to be controlled: first, is the subject sincere? At least, each subject was alone to label, was ensured about confidentiality, has no link with any experimenter, and was explained to the importance of his sincerity. The subject has labeled, on the basis on the video tape, and consequently, it cannot be known if the label concerns the emotional states or the expressions. Second, the labels are mainly given by using the language (French) and some complementary signs. This ability to be comfortable with language to describe self emotions has been several times described (Junca des Morais).

This "naïve" annotation is way to avoid a theoretical decision about emotion characterization and is a good "money exchanges" to give to naïve listeners in further perceptive tests, but of sure they could be insufficient, difficult to compare between speakers and so on.

In fact, the analysis of the labels given by the subjects shown that they use very precise words and signs that allow to "semantic" clear decoding. They annotate expressions, thus different subjects have used "nothing" in the parts where they judged not to have expressed anything. They said after labeling that it does not mean that they did feel "nothing" but just that they did express "nothing". This "nothing" expression has been evaluated in perception tests (Rilliard et al, 2004). They annotate not only primary and secondary emotions but much more general feeling and mental states: "stressed", "anxious", "calm", "surprise", "happy", "disappointed", "worried", "don't understand", "thinking", "concentrated", "confident", "shamed", "perplex", "try to understand", "try different reasoning in my mind", "try to do my best", "bored", "hesitating", "exasperated", "amused"… They annotate mainly with a simple label, often using some quantifiers to label the degree (" + or - " "+/-", "very" , "big", "much", "+", "++", "!!", "little"..). The combination of labels is used to combine different kinds of feeling. They appear quite in the same parts of the scenario for the subjects. For example emotion and thinking: "surprise about results but I don't care", "exasperated but a piece of concentration, I want to do my best", "anxious – little afraid", "complete misunderstanding, lost, exasperated", "disappointed-surprised by results, don't understand, resignated, tired"… Some labels are added by over comments: "oppressed (I am a stressed person)", "worried (again)", "nothing : everything is ok". They can qualify different time layers of the interaction: for example " a little re-ensured, a lot worried".

CONCLUSION

We think that the expert annotation method and the auto-annotation method for emotional labeling have positive and negative qualities in complementary points. When the labels are decisive and must be very fine for further analysis, these labels could be verified by cumulating these two methods. Moreover, these two approaches share the same verifying phases, that is the "naïve" agreement by perception experiments generally based on closed choices.

REFERENCES

1. V. Aubergé, N. Audibert, and A. Rilliard. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, 179-182, 2004.
2. Campbell N (2004) Speech & Expression; the value of a longitudinal corpus, *4th LREC*.
3. Devillers L., Abrillian S. & Martin JC (2005) Representing Real-Life Emotions in Audio-visual Data with non basic Emotional Patterns and context features, *Proc of ACII'05, 519-526.*
4. Gentzler A. L. & Kerns K. A., (2006) Adult attachment and memory of emotional reactions to negative and positive events, *Cognition and Emotion Review, 20 (1), 20-42.*
5. Ghjsen M, Heylen D., Nijholt A., op den Akler R (2005) Facial Affect Displays during Tutorial Sessions.
6. Healy H. & Williams M.G., Autobiographical memory, *Handbook of Cognition and Emotion, Dalgleish & Power eds, 229-242.*
7. Wallbott H & Scherer K (1989), Assessing emotion by questionnaire, Emotion, *4, 55-82, Academic Pres Pub.*

Expressions outside the talk turn: ethograms of the feeling of thinking

Fanny Loyau and Véronique Aubergé

Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
Fanny.Loyau@icp.inpg.fr

ABSTRACT

During our collect of an expressive corpus, a large quantity of non verbal information has also been recorded: chest, facial movements and voice events. We are particularly interested by these actions which happen outside the talk turn, when the subject thinks, and feels about what he thinks. We want to know if these events are real indications of signals about the mental states or the affective states of the subject. In this aim, a typical ethogram methodology has been applied to label these non speech parts into primitive icons of chest movements, facial movements and voice events, not in order to take any decision about the interpretation of what could be expressed by these events, but to classify variant movements into minimal icons.

Keywords

Ethogram, non verbal information, talk turn, feeling of thinking.

INTRODUCTION

In the field of the verbal interaction and expressive communication, many studies are devoted to the expressions transmitted by the speaker during his talk. A few are about the information expressed by the human (or the virtual agent) outside his talk turn, in particular when the subject follows, on line, his interlocutor who is unfolding his talk turn, like "feedback" information ([3] [5]). These information are about his attention, the state of his mental processing - his understanding - on what he gets from the interlocutor, his opinions on what he is decoding, the attitudes and emotions involved by the interaction. Apart his talk turn, the subject can also be in a situation of treatment of a cognitive and/or physical task to carry out. He makes or he lets appear information on his mental and emotional states in the treatment of this task. This kind of situation is characteristic of human-machine interactions.

This paper presents a preliminary analysis of multi-modal expressions of a multi-speaker expressive corpus (Sound Teacher of E-Wiz [1]) in parts of interaction where subjects are not in their talk turns, and in which, however, expressions in voice, face or body are so many and varied. Most of these expressions not have been yet described in litteracy studies (especially not by FACS)We propose here the broad lines of an annotations method of these expressions. This method is not based on an automatic measurement of picture or vocal signals. It restricts role of the human expert to decision of gestural or vocal minimal icons, without interpretation of informative content, like a traditional step of ethogram.

THE EXPRESSIVE CORPUS SOUND TEACHER / EWIZ

The expressive corpus Sound Teacher of E-Wiz [1] was realized from the recording of 17 subjects, 11 women and 6 men. They are placed in a learning situation of vowels of languages of the world using a supposed revolutionary system, Sound Teacher. These subjects are "trapped" by a wizard of Oz scenario: the subject thought to communicate with a computer, whereas the apparent behavior of application is managed remotely by the wizard.

Scenario unfolds into three phases. The first, of training, familiarizes and reassures the subject. The second phase implies the subject in very simple tasks on which he is congratulated, which induced for all subjects overall positive emotions. The third phase is more and more complex, negative judgements are returned to the subject, and ends in a repetition of the initial task, but giving to subjects (false) very bad results which either strongly worried them, or destabilized. After each recording, subjects auto-annotated their production. They noted down, according to their own choice (language, drawings, signs etc) their mental and emotional states finely during the advance of the scenario. Subjects interact only with speech, for the answers or the phases of free comments (neither keyboard nor mouse). They are insulated in soundproof room in front of a screen, and do not know they are recorded. The machine communicates either by text, or by execution of the requested task. Subjects are thus alternatively in phase of reading, thought and production of speech. This is a verbal proposal of a realization of the task, in form of a monosyllabic word.

LABELLING OF CORPUS: A ETHOGRAM

From "feeling of knowing" to "feeling of thinking"

Sound Teacher is a minimal situation of dialogue, since the subject knows that his talk turns do not change the nature of the interaction. The phase of human or humanized communication in which the listening subject sends a feedback to his interlocutor "intentionally" is thus not awaited. However, we further show it, during the "talk turn" of the machine (reading), the subject displays rich expressions during the dynamics of the reading. Especially, during the phase of preparation of his verbal proposal, the subject expresses at the same time mental affects and states. In a task even more specific that of Sound Teacher (A subject is seen raising a question of general culture and does not manage to provide the good answer. He however knows that he knows this answer, he stored it in his memory, and will be able to find it later, but in the present time it is not available.), expressions revealing the mnesic process of the subject was observed, studied and grouped like feeling of knowing [6]. The Sound Teacher task reveals cognitive and emotional processes broader than the mnesic task expressed in feeling of knowing. These processes will be gathered here in a generic phenomenology than we will call "feeling of thinking", the expressions of emotional and mental states.

Methodology

The crucial problem in this study is annotation of expressions. The scenario is known; the "mental and emotional states" are labelled by the subjects (and for some checked in perception experiments). So the subjectivity of labelling by a human "expert" is all the more large. Thus we made choice that ours experts (two experts for 17 subjects) are not informed a priori annotations of the subjects. The goal is to use their competences of human communicating to release a minimal iconicity of the signals. We want to minimize their interpretative competence (they should not be a participating human addition with interaction, but preserve an "objectifying" distance). They must as much as possible bring back to a "iconic syntax" labelling of the movements and vocal events. They must not make "semantic" interpretation of the expression (for example, no labelling of the facial gestures in smiling or others pouts, but geometry and dynamics icons considered to be different). This step, anchored in a methodology of ethogram, is thus fundamentally determined by the minimal icons define as being the labels to be put on the corpus. An additional difficulty is introduced by the non generics of some icons which we were lead to describe, without being able to decide a priori if it is an alternative of a generic icon (i.e. divided by all the subjects, likely to be a communicative signal) or idiosyncratic (i.e. specific to a subject but nevertheless "recovered" indications of communication).

An ethological step

To do it, we have to apply a protocol stemming from ethology (study of customs and individual and social behavior of domestic and wild animals). We chose to annotate ours corpuses with ethograms. Such an object represents the inventory of a species behavior. More exactly, "the ethogram consists of a list of acts and postures observed and defined in a precise way by the experimenter; the scale of observation is built according to this ethogram and allows to quantify the frequency of the behavior over a giving period of time with, possibly, their durations and sequences. Every title is defined according to criteria of direction, sense, localization, distance, intensity or amplitude". So, we can label ours parts of corpuses without speech by using primitive icons for the movements from the chest, those of the face, and the vocal events, without having to make decision which would be of the level of the interpretation. For example, mouth movement presented in figure 1 will not be labelled as "smile" but just as an icon: "lips corner raised", which we call IGS, with as variables intensity, duration and opening or not of the mouth.

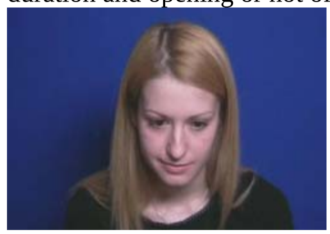


Figure 1: IGS "lips corner raised - right dissymmetry / weak / fast / closed".

Figure 2 presents an extract of our corpus labelling, with the multimedia annotator ELAN (EUDICO (European Distributed Corpora Project) Linguistics Annotator, Max Plank Institute for Psycholinguistics, The Netherlands). There are the speech signal and the video, which are synchronized. On the left there are icons, which are defined in our ethogram (figure 3).

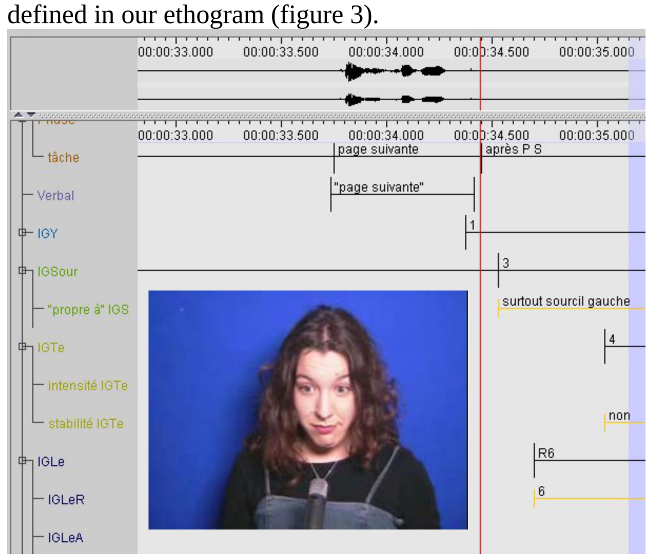
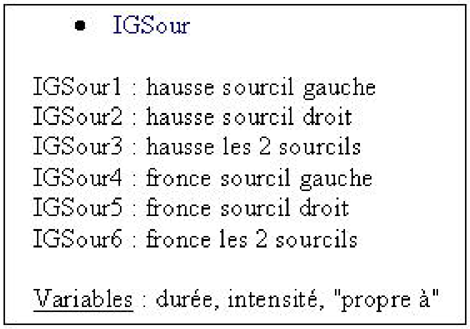


Figure 2: Our labelling, with editor ELAN.

Signals vs. Indicate

Some icons seem to appear in all corpuses, so it would be signals, whereas others seem specific to a subject, it would be indications. But no decision is taken a priori on the various events (movement, face, voice), which will later be identified as being either communication signal, or biological, idiosyncratic indications whose variability will be associated to changes of emotional states [2].



- IGSour

IGSour1 : hausse sourcil gauche
IGSour2 : hausse sourcil droit
IGSour3 : hausse les 2 sourcils
IGSour4 : fronce sourcil gauche
IGSour5 : fronce sourcil droit
IGSour6 : fronce les 2 sourcils

Variables : durée, intensité, "propre à"

Figure 3: A part of our ethogram.

FIRST RESULTS

Temporal micro and macro organizations are fundamental, either when they are coherent with the evolution of emotional states of subjects, or because they reveal these states (some icons will not directly carry information, but temporal organization of these icons will do that [2]).

Temporal organisation

In average, every subject records a 40 minutes corpus (figure 4). The time assigned at speech moments is approximately 8 minutes, thus there is 80 % of communication time which is outside the talk turn. Especially during the 32 minutes used by subject to read the instructions (8 minutes) and to think to his answers (24 minutes), we do not have to find "neutral" position for the chest, the face, nor of completely quiet moment. Something always happened, and it is what we tried to label, in the most objective way possible.

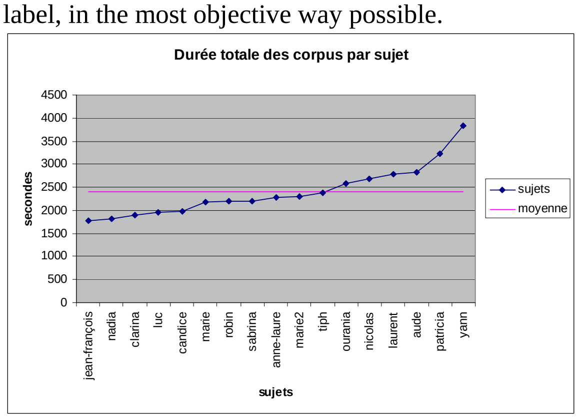


Figure 4: Average time of 17 corpuses.

General characteristics as response times were treated for all subjects, but for the moment the detailed analysis presented below was acieved for 5 of them.

Response time

The average time of answer, corresponding to the time between the moment when the subject hears the stimuli and when he gives his answer is 4,5 seconds. This duration is more important for the training phase, decreases in the following positive phase, then is lengthened again in the last phase, gathering negative induction and destabilization (figure 5).

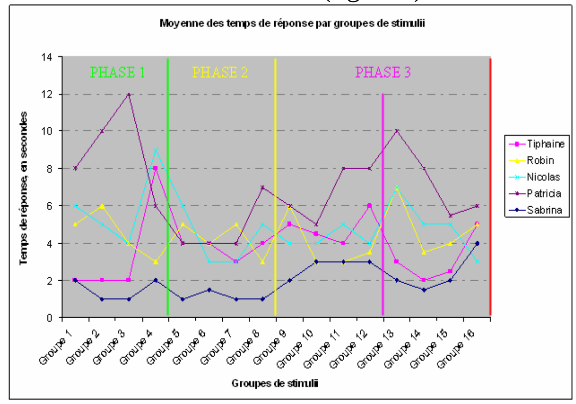


Figure 5 : average of response time, by phases.

The standard deviation are clearly more important for the training and negative phases than for the positive phase, rather stable.

Expressions of "backchannel"

We are particularly interested in the link between occurrences of the various indications outside of the talk turns and the auto-annotation made by subjects, and phases of the scenario.

Common events to all subjects of E-Wiz

Here are the movements found for at all the subjects of E-Wiz: IGSour6 “wrinkle the eyebrows”, IGSour3 “raise the eyebrows”, IGY2 “fold the eyes”, IGRe “glances out of the screen”, IGLp “pass his lips one on the otherIGN1 “fold the nose”, IGLm “bite his lip”, IGMe1 "fold the chin".

Negative situation

Glances out of the screen: Sometimes subjects look outside the screen, in all the phases, but especially when they prepare their answer, and more often in the negative phase. The moments when these glances take place are generally labelled by the subjects themselves as: “perplexity”, “doubt”, “stress”, “trouble”, “approximation”, “perplexity”, “incomprehension”, “irritation”... The glances happening during the talk turns corroborate these results: they happened more frequently during the last phases, and subjects give the same negative label. The laughter: as opposed to what we could have believed a priori, laughter appears more in negative situation than in positive situation. They correspond then to various labels like “tired and amused”, “irritated, anxious”, “stressed, I do not understand”, “surprised, nervous”, “doubt, very aggravated, laughed at my bad performance”, “laughter = tentative of relaxation”, “disappointment but amuses me”, “randomly, a desire for laughing”. The laughter seems here to allow the speaker to change his state, not to remain on an unpleasant negative situation. The majority of the laughter takes place however during the talk turns, but also at the time of negative situations.

The lips protrusion: this indication, if present in all the phases, is also more frequent in the third. It also corresponds to negative auto-annotations: “I look like disappointed”, “concentrate”, “aggravated”, “concentration, boredom”. They never happened at the talk turns time.

Mouth noises: These noises, like whistles, fricatives, plosives, are more frequent and more irregular in the negative phase, and increase particularly in the destabilization part while become also increasingly irregular. Just a few take place at the time of the talk turns.

These behaviors are coherent with the more general results concerning the biological indications of the behavior observed in the players of tennis at the time of situations uncomfortable, of distress [2].

Positive situation

It seems, for the moment, that very few indications are specific to this situation. However, the two first phases, according to the scenario, are supposed to be positive. But in the auto-annotation made by the subjects, few labels are positive, even in these two phases. The most frequent terms are: “boredom”, “irritation”, and “doubt”. The term of “concentration” is also often used, by all the subjects. But according to other words which it is connected it will have a positive connotation (in the first phases), or negative (in the last): “more serious, more concentration”, “great concentration the goal being to understand what is pronounced” / ”boredom, concentration”, “concentration – irritation”.

Smiles: they are more frequent in the two first phases, and agree with the following labels: “quiet”, “enough quiet”, “concentrated”, “proud, pleased”, “astonished and doubts”, “very proud, pleased, astonished”. Thus they are used at very different times as compared to laughter. During the talk turns, that strongly differs from a subject to another. For one there is much laughter during the first phases, for another one there is more out of the talk turns, and this for each phase. For two other subjects there is almost nothing: two in the second phase only for one of these subjects. Finally, for the fifth subject, smiles during the talk turn are not the majority and happen mainly during the last phases.

Specific to the subject

The icon “tilt one’s head” is specific to only one subject, who makes this movement mainly before answering except a stimulus in the destabilization phase, and very mainly on his right (figure 6).

Figure 6: icon IGT "leaning head on the right-hand side".

This icon, which appears in all phases, appears in positive situations, labelled by the subject as “quiet”, “enough quiet”, “concentration”, but especially in negative moments, labelled like “ill at ease”, “oppressed”, “anxiety”...

Another subject is the only one to sniff, especially before giving his answer, and only in the training phase. In auto-annotation it corresponds to parts where this subject feels “concentrated”, or “concentrated, stressed”.

CONCLUSION

In this preliminary work, we tried to outline a still empirical methodology of expressive behaviors ethograms of the subjects in interaction but out of their talk turn. We already could observe that the temporal organization plays a fundamental part, either at the local or the global level. We will confront the identified icons with a statistical validation of their relevance, and with their perceptive validation, at the same time in perception tests of isolated icons, or isolated rhythmic diagrams, and in synthesis with the conversational agent “GRETA” [4].

REFERENCES

1. V. Aubergé, N. Audibert, and A. Rilliard. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, 179-182, 2004.
2. G, Carlier, and C. Graff, to be published. Unpredictability as a counter strategy: An analysis of elite matches*. Journal of Sciences*, 2006.
3. C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behavior. *IVA’05 International Working Conference on Intelligent Virtual Agents*, 2005.
4. I. Poggi, C. Pelachaud, F. de Rosis, V. Caroglio, B. de Carolis. GRETA. A Believable Embodied Conversational Agent. *Multimodal Intelligent Information Presentation*, O. Stock and M. Zancarano, eds, Kluwer, to appear, 2005.
5. M. Schröder, D. Heylen and I. Poggi, to be published. Perception of non-verbal emotional listener feedback. *Speech Prosody 2006*.

Multilevel features for annotating application-driven spontaneous speech corpora

Maffiolo V., Damnati G., Botherel V., Guimier de Neef E. and Maillebuau E.

France Telecom R&D, Technologies Division
Technopole Anticipa, 2 av. Pierre Marzin, 22307 Lannion, France
valerie.maffiolo@francetelecom.com

ABSTRACT

This paper deals with a multilevel annotation of spontaneous speech corpora. Three databases are analysed: recordings of interactions between users and an automatic directory inquiries voice service, recordings of interactions between users and an automatic stock exchange service through a mixed-initiative dialogue system, and recordings of voice messages relating satisfaction regarding the interaction with a customer call centre. In order to identify peculiarities of the spontaneous speech that may be related to the speakers' expressions of emotions, a methodology of annotation is proposed allowing an emotional labelling and a simultaneous annotation, according to prosodic, linguistic, discourse and dialogic indicators, of portions of speech precisely located within the textual transcribing.

Keywords

Spontaneous speech, human-machine interaction, emotion, annotation.

CONTEXT

Motivated by the development of human machine interfaces that are more adaptive and responsive to a user's behaviour, we aim at building a detection and characterization model of expressions of emotions in speech. Among topics of interest, there are the content processing for classification of voice messages and the improvement of both naturalness and efficiency of spoken language human machine interfaces (Cowie et al., 2001).

Several challenges are involved with developing systems to detect expressions of emotion in voice. First, data collection of real-life scenarios is often difficult to acquire, limiting much of the research on emotion to use experienced speakers (actors) to simulate a specific emotional state (Polzin and Waibel, 1998; Yacoub et al., 2003) or to control the user-system interaction by wizard-of-oz in order to give rise to expressions of emotion (Ang et al., 2002; Batliner et al., 2003). Moreover, vast amounts of data characterizing various emotions types and from a large number of users and contexts are needed to design valid detection models.

Second, the description of the emotional states expressed in speech, including the definition of emotion categories and the location of emotional utterance in speech (determining which portion of a speaker's utterance should be taken as emotional and which as neutral) is a complex problem mainly due to the lack of consistency in the definitions of emotions (Cowie and Cornelius, 2003; Scherer, 2003; Vidrascu and Devillers, 2005).

Third, identification of relevant peculiarities of spontaneous speech correlated to expressions of emotions is of prime importance. Although the main research in the area of emotional speech is based on extraction of acoustic features, more and more works dealing with conversational interfaces attempt to combine prosodic features with other sources, as lexical, semantic and discourse information, to maximize emotion detection (Batliner et al., 2003; Lee and Narayanan, 2005; Liscombe et al., 2005; Litman and Forbes-Riley, 2005).

In this paper, we present the databases analysed and the coding schemes used. Our aims of improving voice messages classification and building a detection model of expressions of emotions in speech in human-machine interaction bring us to collect a large amount of data and to analyse them with an emotional labelling and a multilevel annotation based on prosodic, linguistic, discourse and dialogic indicators.

DATABASES

In our study, the corpora stem from experimentations based on real-life human-machine interactions. They were obtained over the telephone from real customers either leaving a voice message on an answering machine or engaged in spoken dialog with automatic systems. To be more precise, three spontaneous speech databases are analysed.

Recordings of interactions between users and an automatic directory inquiries voice service

The elocution is constrained and the interaction is based on isolated word recognition. The dialogue's design induces the system to formulate proposals that the user can accept or reject. If these proposals are rejected, there may be expressions of negative emotions by the user, which must be detected to adapt the dialogue. The corpus contains 1666 dialogues (39.7% of male speakers and 60.3% of female speakers) representing 9670 utterances.

Recordings of interactions between users and an automatic stock exchange service through a natural and mixed-initiative dialogue system

The objective of the application (stock exchange transactions) is likely to induce emotions. Users speak without language constraints. So, there are potentially more linguistic peculiarities than within the more directed dialogue with the automatic directory service. The corpus contains 217 dialogues (77% of male speakers and 23% of female speakers) representing 5406 utterances.

Recordings of voice messages relating satisfaction regarding the interaction with a customer call centre

Users are invited through a short message to call a toll-free number where they can express their satisfaction with regards to the customer service they recently called. Calling this toll-free number, they are asked to leave a message, after hearing the following prompt: "*[...] You recently contacted our customer service. Thank you for calling us. We would like to make sure that you've been satisfied with the reception and with the way your request was followed up. Don't hesitate to make any comment or suggestion about our service that can help us make it better [...]*"
Studying this database from an emotional point of view is interesting to improve the automatic classification of messages. The particularity lies in the potentially long length of the messages, the highly spontaneous expressions (repairs, disfluencies…) and the fact that several different expressions of emotions can occur in a same message. The corpus contains 1778 messages.

ANNOTATION SCHEMES

The aim of the annotation is to identify peculiarities of the spontaneous speech that may be related to the speakers' expressions of emotions. An automatic annotation (mainly with acoustic indicators) and a manual annotation have been put in place. In this paper, only the manual annotation is discussed. Both the voice messages and dialogues between users and automatic systems are textually transcribed and annotated by an expert, with dedicated software. Annotation is done from the listening of recordings and the tags are inserted within the exact transcribing of the spontaneous speech. The coding scheme for voice messages is based on prosodic and linguistic indicators. The coding scheme for spoken dialogues with automatic services is based on prosodic, linguistic, discourse and dialogic indicators. These latter partly take into account the context of the interaction.

Prosodic indicators

The prosodic indicators concern hyper-articulation, rhythm variation, loudness variation, breaking-up, vowel lengthening, silent pauses and intelligibility pauses.

Linguistic indicators

The linguistic first-level indicators concern disfluency indicators (discourse marker, repetition, false start...).

The prosodic and linguistic indicators are seen as intra-utterance indicators, whereas discourse and dialogic indicators are seen as inter-utterance indicators. These latter are tagged at the utterance level in order to annotate the dialogue's course.

Discourse indicators

The discourse indicators distinguish private conversation and comment. Private conversation is a conversation between the user and another person, the purpose of their conversation has no relation with the interaction with the service or the purpose of the call (in the case of voice message). Users' comments are related to the interaction with the service. Often users talk to themselves.

Dialogic indicators

Concerning the dialogic indicators, repetitions and reformulations are potentially indicators of changes in a speaker's attitude because they often occur in the case of a misunderstanding of the system. They constitute events which may potentially trigger off emotions. We distinguish between two ways of repeating the content of the former utterance: repetitions in which the same utterance is repeated and reformulations in which the user chooses different words to convey the same content.

Emotional labelling

The emotional labelling of the data is done independently of the peculiarities annotation. A compromise has been found between the precision of the description of perceived emotional states and the simplicity of labelling, leading to four mutually exclusive emotional labels: “positive”; “negative”; “very negative” and the default “neutral”. Emotional labels are attributed to portions of speech. The tags encompass words sequences. This choice is justified by the fact that particularly voice messages may contain expressions of different emotions.

OBSERVATIONS ON ANNOTATED CORPORA

Man/machine interactions

Concerning the spoken dialogues with automatic services, most utterances are neutral in nature. They have no apparent display of emotions. Indeed, although 39% of the stock exchange dialogues contain expressions of negative emotions, only 8.3% utterances contain negative emotion comparing to 90.8% of neutral. Identically, although 35% of the automatic directory dialogues contain expressions of negative emotions, only 10,5% utterances contain negative emotion comparing to 83.2% of neutral (Table 1).
Regarding specifically the data from automatic directory service, it appears that negative emotions are mainly expressed by users when the interaction with the service is less efficient than expected by users. Few emotions are expressed in the first speech turns in the dialogues. However, the more the dialogues are long, the more negative emotions are likely to be expressed.

	Stock Exchange (%)	Automatic Directory (%)
Negative	7,7	10
Neutral	90,8	83,2
Positive	0,9	6,3
Very Negative	0,6	0,5

Table 2. Distribution of the emotional labels for the two automatic system-based interaction databases

Voice messages

From the overall set of 1778 messages that has been fully annotated, 847 messages (47,6%) have been labeled with at least one portion of emotion expression. Among these, 799 contain just one type of label (positive OR negative OR very negative) and 48 contain two different types of labels (for example one "negative" portion and one "positive" portion but also possibly one "negative" portion and two "positive" portions).

	Positive	Negative	Very Negative
Nb. portions	585	436	43
Nb. different messages	511	351	33
Nb. occurrences of words	7578	11683	1435
Nb. different words	790	1593	456
Average nb. of words per portion	12,9	26,8	33,4

Table 3. Description of the emotional labels for the voice messages database

From Table 2, a first observation can be easily made about the length of the negative portions. On average, negative portions are longer than positive ones. What's more, the lexical variety is more important for negative messages.
Due to the lack of constraints on users' elocution and to the nature of the open question users are submitted to, messages are characterized by a high level of disfluency.
The Table 3 illustrates how the linguistic indicators are distributed among the emotional portions of messages.
Negative and very negative segments distinguish from the positive ones through a higher level of disfluency. This could be only due to the difference observed in terms of average number of words per segments, but other ratios evaluated at the word level confirmed this tendency.

	Positive	Negative	Very Negative
False starts	6,5 %	16,7 %	13,9 %
Discourse markers	24,8 %	33,0 %	27,9 %
Re-starts	15,0 %	28,0 %	27,9 %
Repetitions	18,1 %	20,1 %	23,3 %
Filled pauses	34,2 %	45,6 %	37,2 %

Table 4. Percentage of emotional portions containing at least one linguistic indicator

HIGHER LEVEL LINGUISTIC LABELLING

In order to go beyond linguistic features described previously, a study has been carried out aiming at determining other linguistic markers. A typology of these markers was drawn up from the observation of the voice messages corpus. For this corpus, markers potentially related to emotion and/or satisfaction are proposed.

Lexical markers

90 lexical markers in keeping with (positive or negative) emotion and satisfaction have been identified, among which:

- slang words or expressions, familiar language, ex: "merde" (*shit*), "ennuyeux" (*annoying*)
- opinion or sentiment vocabulary, ex: "heureux" (*happy*), "satisfait" (*satisfied*)
- vocabulary connoting satisfaction or emotion with respect to the context, ex: "problème" (*problem*)

These markers can be divided into 2 lists: strong markers that are sufficient to detect emotion or satisfaction, and weak markers which need the taking into account of the context (Table 4). For instance, in the voice messages corpus, verbs in keeping with the service may convey satisfaction or dissatisfaction according to the context:
Ex: "renseigné" (*informed*) + "mal" (*ill, bad*)
= "mal renseigné" (*ill-informed*) => dissatisfaction

	Satisfaction	Emotion
Strong markers	61%	17%
Weak markers	20%	2%

Table 4. Distribution of the satisfaction and emotion-related lexical markers

Syntactic markers

A shallow analysis of the syntactic context may be sufficient either to qualify the polarity of the emotion:

- a negative context reverse the polarity of the satisfaction or the emotion, ex: "soucis" (*worry*) = dissatisfaction / "sans soucis" (*free from worry*) = satisfaction

Or to detect expression of emotion:

- imperative forms of verbs carry out dissatisfaction, ex: "remboursez-moi" (*refund me*)
- as well as passive forms of some verbs, ex: "j'ai pas été renseigné" (*I haven't been informed*)

About 80 syntactic contexts involving noun phrases, verb phrases, prepositional phrase, adjectival phrase and expressions have been formalized to characterize expressions of emotion and/or satisfaction.

Semantic and stylistic markers

In many cases, emotion and satisfaction are connoted in the corpus through semantic and stylistic markers:

- speech acts: intimidations, orders, complaints, encouragements
- stylistic markers: accumulation of negative facts connoting dissatisfaction
- pragmatic markers: reiteration of negative facts

DISCUSSION

The strong points of this study are on the one hand to deal with real spontaneous speech corpora. It is a strong point even if their collection is not specifically dedicated to expressions of emotions and even if the emotions collected are not very various (negative emotions are mainly anger and frustration, positive emotions are mainly relief and delight) and not very expressed. On the other hand, the methodology of annotation used allows both an analysis of portions of speech precisely located within the textual transcribing and a simultaneous annotation according to a range of indicators allowing the combination of different sources of information. This methodology induces a very precise and fine annotation. We have opted in a first step for a complete annotation of our databases with only one expert-coder. Although the intra-coder coherence has been checked, the weak point of our study is the lack of validation of the annotation by several coders. To overcome this point, we aim at having more coders to evaluate agreement among coders on each of the three databases.

The span of the manual annotation has been chosen quite large on purpose, in order to study the correlation of the various indicators with the expressions of emotions. Further work will focus on the automatic detection of features that will be identified as relevant indicators and the integration of a real-time emotion detection model with automatic speech recognition.

Concerning perspectives on the automatic identification of high-level linguistic markers, first experiments have shown that strong lexical or syntactic markers can be detected by a parser with dedicated resources. On the other hand, the identification of semantic, pragmatic, and stylistic markers will necessitate a high-level linguistic analysis.

ACKNOWLEDGMENTS

The authors thank I. Sari for having annotated the databases and C. Blouin and V. Gasser for having participated to the elaboration of coding schemes.

REFERENCES

1. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J.G.: Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* (2001), 33-80.
2. Polzin, T.S., and Waibel, A.H.: Detecting emotions in speech. In *Proc. Cooperative Multimodal Communication* (1998).
3. Yacoub, S., Simske, S., Lin, X., and Burns, J.: Recognition of Emotions in Interactive Voice Response Systems. In *Proc. Eurospeech 2003* (2003), 729-732, Geneva.
4. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. Intl. Conf. on Spoken Language Processing, 3* (2002), 2037-2040, Denver.
5. Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E.: How to find trouble in communication. *Speech Communication 40* (2003), 117-143.
6. Cowie, R., and Cornelius, R.R.: Describing the emotional states that are expressed in speech. *Speech Communication 40* (2003), 5-32.
7. Scherer, K.R.: Vocal communication of emotion: a review of research paradigms. *Speech Communication 40* (2003), 227-256.
8. Vidrascu, L., and Devillers, L.: Real-life Emotion Representation and Detection in Call Centers Data. In *Proc. Affective Computing and Intelligent Interaction* (2005), Beijing.
9. Lee, C.M., and Narayanan, S.S.: Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing 13, 2* (2005), 293-303.
10. Liscombe, J., Riccardi,G., and Hakkani-Tür, D.: Using Context to Improve Emotion Detection in Spoken Dialog Systems. In *Proc. Interspeech* (2005), Lisboa.
11. Litman, D.J., and Forbes-Riley, K.: Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, in press (2005).

Music as a Method of Identifying Emotional Speech

Maartje Schreuder
University of Groningen
Department of Linguistics
P.O.Box 716
9700 AS Groningen
The Netherlands
M.J.Schreuder@rug.nl

Laura van Eerten
University of Groningen
Department of Linguistics
P.O.Box 716
9700 AS Groningen
The Netherlands
Laura.van.Eerten@gmail.com

Dicky Gilbers
University of Groningen
Department of Linguistics
P.O.Box 716
9700 AS Groningen
The Netherlands
D.G.Gilbers@let.rug.nl

ABSTRACT

In this pilot study we investigate whether differences in emotional speech are characterized by musical modalities. In music sad and cheerful melodies are often distinguished, respectively, by a minor and a major key. Our aim is the identification of analogous interval differences in the pitch contours of emotional speech. We recorded and analyzed the performances of professional readers reading passages from A.A. Milne's Winnie the Pooh in Dutch. We are interested in the sad character Eeyore and the happy, energetic Tigger. Although we do not find modality in the pitch contours of all speakers, we do find intervals between tones indicating minor modality exclusively in Eeyore passages and intervals indicating major modality exclusively in Tigger passages.

Keywords

Laboratory Phonology, Musicology, Emotional Intonation.

INTRODUCTION

Composer Fred Lerdahl and linguist Ray Jackendoff point out the resemblance between the ways both linguists and musicologists structure their research objects [15]. This insight gave rise to the proposal of a formal generative theory of tonal music [16], in which they describe musical intuition. Above all, insights from non-linear phonology [17,18,23,12] led to scores provided with tree structures, indicating heads and dependent constituents in the investigated domains. In this way, Lerdahl and Jackendoff bring to life a synthesis of linguistic methodology and the insights of music theory. [8] Shows that music theory in turn can be useful to describe linguistic rhythmic variability. Further examples of musical and linguistic cross-pollination are [11,17,2,20,13,14,9]. These studies provide arguments for the proposition that every form of temporally ordered behaviour, like language and music, is structured the same way. In both disciplines the research object is structured hierarchically and in each domain the important and less important constituents are defined, which enables the listener to interpret the stream of sounds.

In the present pilot study, we investigate whether the similarities between music and language can be extended to the area of extra-linguistic characteristics, such as emotion. Therefore, we compare intonation patterns in speech to musical melodies.

THEORETICAL BACKGROUND

The scale in western tonal music is divided into twelve steps, also called 'semitones'. Typical for the minor modality is that it features chords that are characterized by a distance of three semitones between the tonic and the (minor) third, whereas chords in the major modality feature a distance of four semitones between the tonic and the (major) third. This difference in thirds is the main factor for the perception of mood in music. Sad and cheerful music is often described as a difference between, respectively, a minor and a major key, although in some instances composers play around with the notions of major and minor modality, which may result in cheerful music in a minor key, or sad music in the major key.

In Figure 1, the keys of a keyboard instrument are shown. The distance between C and C#, for instance, involves one semitone; the distance between C and D two semitones. Thus, a minor third is constituted by C and E*b* and a major third by C and E.

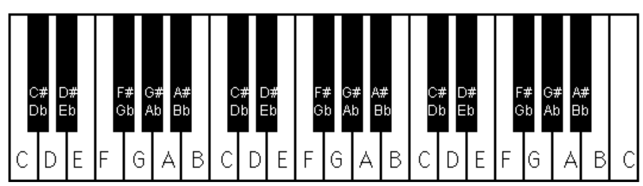


Figure 1. Keyboard.

Each note has a corresponding frequency. For example, the concert A is 440 Hz. A' one octave higher has a double frequency: 880 Hz; A one octave lower has a frequency of 220 Hz. Within the octave, A and A' are twelve semitones apart: five black keys and seven white keys in Figure 1. The frequency ratio between two semitones is equal. It is the twelfth root of two, which is approximately 1.0595. Table 1 shows frequency values of each note.

[4] Studied Dutch speech and found out that the majority of the speakers speak according to an internal tuned scale. [5, 6] Investigate the modality of Japanese emotional speech. Normally, the pitch range of seven or more semitones is used in sentences. [6] Conclude that utterances perceived as having positive affect significantly show major-like pitch structure, whereas sentences with negative affect have a tendency to minor-like pitch structure. The conclusions are based on cluster analyses of the pitch contours of recorded utterances. In these cluster analyses the actual pitch values at every millisecond are rounded off to the value of the nearest semitone (*cf.* Table 1). The result is a graph in which one can read which semitones occur most in the utterance.
In this pilot study, we present a follow-up to these studies in which we try to find out whether Dutch emotional speech can be identified by musical modalities.

Note	Freq.	Note	Freq.	Note	Freq.
C	65.4 Hz	C	130.8 Hz	C	261.6 Hz
C#	69.3 Hz	C#	138.6 Hz	C#	277.2 Hz
D	73.4 Hz	D	146.8 Hz	D	293.6 Hz
D#	77.8 Hz	D#	155.6 Hz	D#	311.2 Hz
E	82.4 Hz	E	164.8 Hz	E	329.6 Hz
F	87.3 Hz	F	174.6 Hz	F	349.2 Hz
F#	92.5 Hz	F#	185.0 Hz	F#	370.0 Hz
G	98.0 Hz	G	196.0 Hz	G	392.0 Hz
G#	103.9 Hz	G#	207.7 Hz	G#	415.3 Hz
A	110.0 Hz	A	220.0 Hz	**A**	**440.0 Hz**
A#	116.6 Hz	A#	233.2 Hz	A#	466.2 Hz
B	123.5 Hz	B	247.0 Hz	B	493.9 Hz

Table 1. Approximate note frequencies in Hz.

The method in [5,6] has as a drawback that it is not clear whether the most frequent notes occur as direct sequences. Therefore, we will also investigate sequences of individual notes in scores of emotional speech apart from cluster analyses.

METHOD

In order to obtain different emotions in speech, we asked five primary school teachers to read out selected passages in Dutch from Winnie the Pooh [19], in which energetic, happy Tigger, and distrustful, sad Eeyore, are presented as talking characters.
The primary school teachers are experienced readers. The two men and three women aged 27 to 32 all claimed to have musical affinity; four of them played an instrument. They all read out the same passages, which were recorded on hard disk as wav-files and analyzed using the software programs CoolEdit 2000 and Praat [3].
The passages in which Tigger and Eeyore speak were extracted and concatenated into ten files each varying from 8 to 53 seconds. The pitch information of these files was measured every ten milliseconds using Praat. In this way we obtained sequences of frequency values representing the pitch contours. Comparison to the original pitch contours revealed a great similarity. Therefore, we decided that this sample rate of ten milliseconds was sufficient for our experiment.
Subsequently, we did a cluster analysis of the pitch data in order to find out which frequencies occurred most in each contour. For this cluster analysis we relied on a cluster algorithm in Excel presented in [5,6,10,22]. The product of the frequency data was calculated, and assigned to the nearest semitone in an equally tempered scale, resulting in a semitone power spectrum. In other words, the obtained pitch values were clustered i.e. rounded down or up to the value of the nearest semitone. This normalization procedure resulted in a semitone histogram in which one can read which semitones occur most in the utterance. In this way, we made an abstraction of the real pitch values that can be compared to the abstractions phonologists make when they describe various allophones as the realizations of one and the same phoneme. As [5] remarks, it might be more valid to normalize to the speaker's dominant pitches above the tonic, instead of to the musical equally tempered scale, and then study the interval substructure. This would probably lead to somewhat different results, but it would also complicate the analyses.
Furthermore, we converted the pitch contours of the stories into musical scores, to account for intervals in sequences. The aspect of time may be an important property in the analyses of modality.

ANALYSES AND RESULTS

Cluster analysis

[6] Identify the musical modality of Japanese speech on three peaks in the cluster analysis, because musical modality is based on triads. [21], however, claim that the range of Dutch intonation moves between two perceptively relevant declination levels in contrast to the three levels of English intonation. Indeed, most of our graphs show one or two peaks. There are only two graphs with three peaks. Therefore, we decided to indicate the modality on the occurrence of intervals of thirds in the graphs. If the interval between peaks concerns a minor third, we indicate the modality of speech as minor; if the interval concerns a major third, the modality is considered to be major.
Inspection of the cluster analyses shows that not all graphs contain more than one peak. In other words, in graphs with just one peak the modality cannot be determined. These one-peak graphs were found in eight of our twenty sound files. In contrast to tonal music, which usually has a major or minor modality, speech can be neutral. Music with a neutral modality also occurs. (Metal) rock music, for instance, frequently uses so-called power chords, which consist solely of the tonic and the dominant. Without triads, no modality can be derived. Moreover, one can think of music without chords, with a melodic line with intervals of e.g. only fourths and fifths. This is a rare phenomenon in music, while it seems to be a normal option in speech.
In five cases in our experiment the peaks are too far apart to decide on the modality. If the peaks constitute a fifth, for example, one cannot determine the modality. This does not immediately imply that all these instances are

counterexamples, they are just indecisive. Seven cases remained for analysis.

Our analyses confirm our hypothesis. The major modality is exclusively found in sound files of Tigger stories in which thirds were observed, whereas the minor modality only appears in sound files of Eeyore stories. We conclude that Tigger speaks in a major key and Eeyore in a minor key.

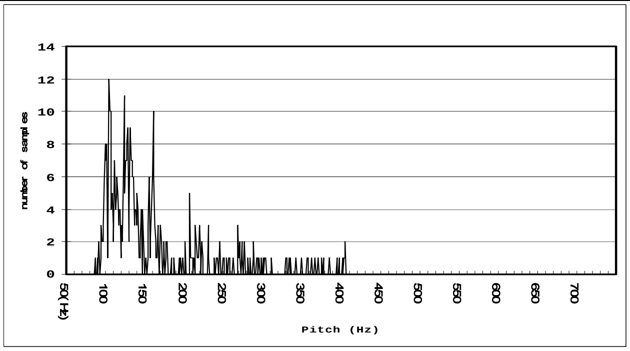


Figure 2a. Tigger in major, cluster analysis.

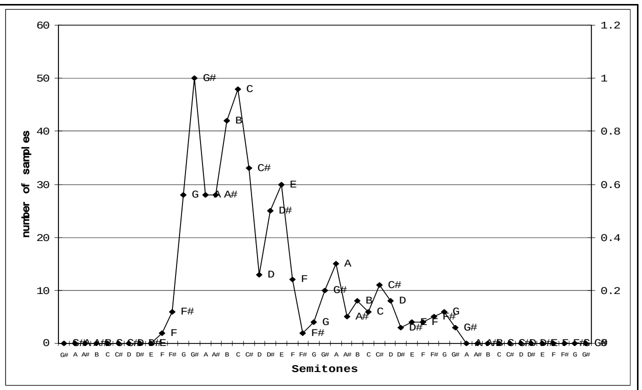


Figure 2b. Tigger in major, semitones.

Figure 2a shows a cluster analysis example of the raw data of Tigger as performed by subject HJ. The x-axis presents the pitch values in Hertz and the y-axis depicts the number of occurrences of a certain pitch value in the sound file. The frequency range is large, from 87 to 406 Hz.

Figure 2b shows the same fragment as Figure 2a, this time clustered in semitones. The figures were obtained using the cluster algorithm macro in Excel [5,6]. On the x-axis abstractions (musical phonemes) of the real frequencies (musical allophones) are depicted as musical notes. On the y-axis we show the number of samples for each note. Our analyses are based on the semitone graphs, such as the one in Figure 2b.

Figure 2b is one of the few graphs that show three peaks. From left to right the first two peaks are on the notes G# and C. The distance of four semitones between these notes constitutes a major third. The following peak in the graph is at the note E which also constitutes a major third with the preceding C. G# and C form an inverted major third together. Tigger, as spoken by the male subject HJ, is a cheerful character and his speech indeed exhibits the major thirds of a major modality.

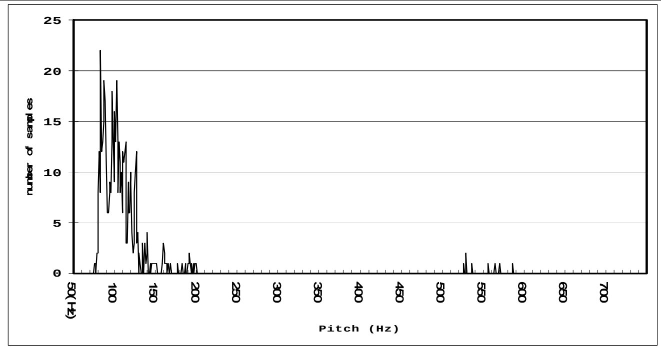


Figure 3a. Eeyore in minor, cluster analysis.

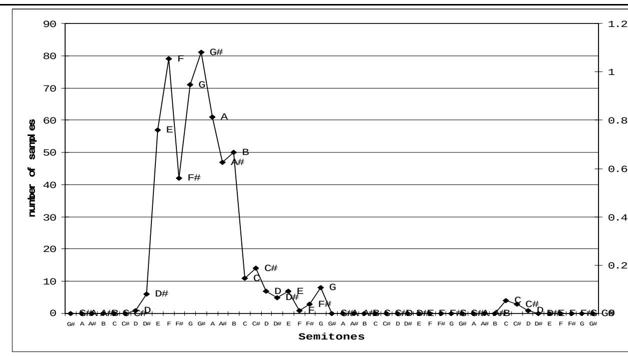


Figure 3b. Eeyore in minor, semitones.

Figure 3a shows the clustered data of the same subject HJ's interpretation of Eeyore. The frequency range is smaller this time, from 75 to 200 Hz. In comparison, the frequency range of Tigger was from 87 to 406 Hz. The peaks are also located in lower regions in comparison with Tigger.

Figure 3b shows the same fragment clustered in semitones with two peaks on, respectively, F and G# (or Ab). The distance between the peaks is three semitones, in other words a minor third: Eeyore speaks in a minor modality.

Musical Scores

The cluster analysis ignores absolute intervals in time. In other words, the result is not a kind of musical score of speech. Actually, we do not know whether peaks on, for instance, C and E constitute a major third or an inverted augmented fifth. [5] Justifies his choice by claiming that it is unlikely that simply an alteration in the sequence of pitches that conveys positive or negative affect could transform a minor mood into major, or vice versa [5]. In music, however, the same melody can cause different moods depending on the chord structure of the song. For example, if a phrase in the key of C is repeated, whilst the chord progression changes to A minor, which is the parallel of C, the mood may change from cheerful to sad.

Therefore, we incorporated time as a factor, which may lead to more reliable results. We did this by using the following formula in Praat: 2 ^ (round (log2 (self/440) * 12) / 12) * 440, which works similarly to a vocoder/harmonizer, rounding off automatically all frequency values at semitone value. The formula calculates the twelfth root of two for rounding off all tones to their nearest semitone, using 440 Hz, the concert A, as a reference tone. Figure 4 shows that, although this manipulation does change the original values, the differences are very small and do not reach a perceptible level.

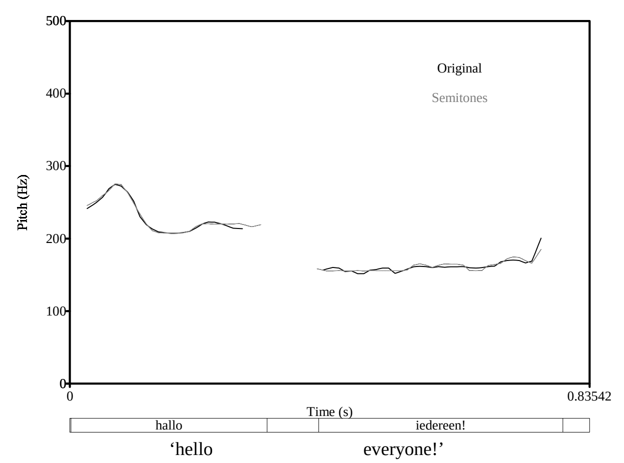


Figure 4. Pitch contour of the original speech sound compared to the contour rounded off to the nearest semitone values (Tigger).

Consequently, the manipulated pitch objects were resampled to sine waves. We converted these sine waves to MIDI files, using the freeware program AmazingMIDI [1]. MIDI files can be represented as musical scores by means of e.g. Steinberg Cubase software or Sibelius. In this way, the resulting musical score of a sound file enables us to determine the key and the modality of the speech.

The resulting scores of two stories, the same stories as depicted in the cluster analyses in Figure 2 and Figure 3, are shown in Figure 5 and Figure 6. These scores are simplified versions, because a pitch contour consists of several 'glissandos', while the MIDI file must sample the tones into distinct notes. We chose to convert the tones into eighth notes, with the result that all notes of one glissando were unified into single chords. From these chords we chose the most prominent note for each syllable sounding in the original pitch contour. For readability reasons, the Tigger score is in the treble clef, while Eeyore spoke in a lower tone region and is therefore set in the bass clef.

Figure 5. Musical score of the same Tigger story as in Figure 2.

In this score of the short Tigger monologue we see the same notes stand out as in Figure 2: G#, C and E, but also A and B. A and B do not form thirds with the other notes. The objective of this score was to look whether (prominent) adjacent notes, ideally notes on neighbouring stressed syllables, form thirds in sequence. This, however, is hard to extract from the score in Figure 5, because most intervals between notes in sequences are larger intervals than thirds. Moreover, most phrases appear to be spoken on a single tone. Comparing intervals between different phrases would be wrong, because in the original speech file parts of text intervened between these phrases. We find some thirds on stressed syllables, however, which appear to be major thirds: the interval G# – E between *lo* and *ie* in *Hallo iedereen* 'hello everyone', and the interval C# – A between *ter* and *Ie* in *achter Iejoor* 'behind Eeyore'. The major part of this score is built upon notes which form major thirds with each other. This gives the ultimate feeling of a major key: a happy, cheerful, and energetic story. Figure 6 gives the score of the Eeyore monologue. Again we see many Fs and A*b*s, as in the cluster analyses in Figure 3.

Figure 6. Musical score of the same Eeyore story as in Figure 3.

The story is longer, and here we are able to identify sequences of thirds between stressed syllables. Examples are G*b* – A in the syllables *maak* and *het* in *hoe maak je het?* ‘how do you do?’, and F – A*b* in the syllables *één* and *an* in *de één of ander* ‘someone or other’.
We did not make (simplified) scores of all the stories. The cluster analyses seem to give a good account of the internal relations in the melodies. While the energetic Tigger speaks in a major key, the melancholic character Eeyore expresses himself in a minor key.

CONCLUSION

In this pilot study we analyzed clustered frequency peaks in stories in which the happy Tigger and the sad Eeyore were speaking characters, and we derived musical scores of the pitch contours. The results show that in the cases in which we do find intervals of thirds between the frequency peaks, the major modality is always observed in sound files of Tigger stories, whereas the minor modality is observed in sound files of Eeyore stories. Although thirds were only found in a minority of our material, there were no counterexamples in the fragments containing thirds. The derived musical scores of the intonation contours show that at least the minor thirds of Eeyore can also be found in sequences of stressed syllables.
Although speech can be neutral, we found a tendency that a sad mood can be expressed by using intervals of three semitones, i.e. minor thirds. Cheerful speech mostly has bigger intervals than thirds, but when thirds are used, these thirds tend to be major thirds. Strong conclusions cannot be drawn from only one such a small-scale experiment using a new analytic technique. But the evidence presented above is certainly suggestive. At the very least, these results are an indication that the mood of emotional prosody in speech is rather similar to musical modality. Therefore, this could be a promising method to study emotion in speech. At least, the tendency we found suggests that further investigation of the similarities between music and speech could be fruitful.

Acknowledgements

We are grateful to Norman Cook, Paul Boersma, the participants of the experiments, and the anonymous reviewers of the abstract.

REFERENCES

1.Amazing MIDI: Araki Software, Japan. http://www.pluto.dti.ne.jp/~araki/amazingmidi/ (1998-2003)

2.Attridge, D. *The rhythms of English poetry*. English series no. 14. Burnt Hill, Essex: Longman (1982).

3.Boersma, P. & D. Weenink. Praat: a system for doing phonetics. http://www.praat.org. (1992-2004).

4.Braun, M. Speech mirrors norm-tones: Absolute pitch as a normal but precognitive trait. *Acoustics research letters on line 2*, 3 (2001), 85-90.

5.Cook, N.D. *Tone of voice and mind. the connections between intonation, emotion, cognition and consciousness*. Amsterdam: John Benjamins (2002).

6.Cook, N.D., T.Fujisawa & K. Takami. Application of a psycho-acoustical model of harmony to speech prosody. *Proceedings of speech prosody*. Nara, Japan (2004), 147-150.

7.Eerten, L.J.A. van. *Mineur en majeur in emotionele spraak, een intonatieonderzoek.* Ms, University of Groningen (2004).

8.Gilbers, D.G. Ritmische Structuur. *Glot 10* (1987), 271-292.

9.Gilbers, D.G. *Phonological Networks: a theory of segment representation*. Phd dissertation. Grodil 3, University of Groningen (1992).

10.Gilbers, D.G. & M.J. Schreuder. *Language and music in optimality theory*. [Rutgers Optimality Archive (2002), 571-0103].

11.Guéron, J. The meter of nursery rhymes: an application of the Halle-Keyser theory of meter. *Poetics 12* (1974), 73-110.

12.Hayes, B. The Phonology of Rhythm in English. *Linguistic Inquiry 15*, 1 (1984), 33-74.

13.Hayes, B. & A. Kaun. The role of phonological phrasing in sung and chanted verse. *The linguistic review 13*, 3-4 (1996), 243-304.

14.Hayes, B. & M. MacEachern. Quatrain form in English folk verse. *Language 74* (1998), 473-507.

15.Jackendoff, R. & F. Lerdahl. *A deep parallel between music and language*, Ms. Indiana University Linguistic Club (1980).

16.Lerdahl, F. & R. Jackendoff. *A generative theory of tonal music*. The MIT Press, Cambridge, Massachusetts, London, England (1983).

17.Liberman, M. *The Intonational System of English*. Garland Publishing, Inc., New York & London (1975).

18.Liberman, M. & A. Prince. On Stress and Linguistic Rhythm. *Linguistic Inquiry 8*, 2 (1977), 249-336.

19.Milne, A.A. *Winnie de Poeh*. Translated by M. Bouhuys. Van Goor, Amsterdam (1994).

20.Oehrle, R. Temporal structures in verse design. In: *P.Kiparsky & G. Youmans (eds) Rhythm and Meter:* San Diego: Academic Press (1989), 87-119.

21.Nooteboom, S.G. & A. Cohen. *Spreken en verstaan*. Van Gorcum, Assen (1995).

22.Schreuder, M.J. *Prosodic processes in speech and music*. PhD dissertation. University of Groningen (to appear).

23.Selkirk, E.O. *Phonology and Syntax: The Relation Between Sound and Structure*. Cambridge, Mass.: MIT Press (1984).

A Framework for Generating and Indexing Induced Emotional Voice Data

Aaron S. Master
Department of Electrical Engineering
Stanford University
Stanford, CA, USA
amaster@gmail.com

Ing-Marie Jonsson
Department of Communication
Stanford University
Stanford, CA, USA
ingmarie@csli.stanford.edu

Clifford Nass
Department of Communication
Stanford University
Stanford, CA, USA
nass@stanford.edu

Peter X. Deng
Stanford University,
Stanford, CA, USA
pdeng@stanford.edu

Kristin Richards
Stanford University,
Stanford, CA, USA
klr10@stanford.edu

ABSTRACT
In the field of automatic voice emotion recognition there are basically two tasks: generating "emotional" voice data and classifying it automatically. We presently focus on the former, which to date has not seen standardization across the literature. We provide a brief overview of methods and practices for generating and labeling emotional voice data, and highlight strengths and weaknesses of the various approaches. Given this context, we argue for an induced emotion corpus, multiply-labeled by emotional condition and user response to a standard affect scale, obtained in an automated fashion. In support of this method, we cite recent pilot studies where contrasting emotional video clips and readings led to significant differences on subjects' responses to standard affect scales; in one of the studies voice recordings were made and significant differences were automatically detected between conditions.

Keywords
Voice emotion, emotion corpora, induced emotion, emotion labeling, PANAS.

INTRODUCTION
Recent years have seen significant strides in voice technology. Automatic speech recognition, convincing synthetic speech, and automatic speaker verification are some better-known examples. Voice emotion recognition, however, has yet to see significant commercial application and in many regards remains in its infancy. The field basically consists of two tasks: generating "emotional" voice data and classifying it automatically. We presently focus on the former, which to date has not seen standardization across the literature. Our goal is to help clarify what type of data is needed, and how it may reliably and efficiently be obtained. We ultimately argue for an induced emotion corpus, double-labeled by emotional condition and user response to a standard affect scale, obtained in an automated fashion.

This paper is organized so that the next section provides a brief overview of methods and practices for generating and labeling emotional voice data. The two sections following sections offer a more in-depth treatment of emotions and labeling of emotional data. In the fifth section, we propose an implementation framework that addresses privacy concerns. We conclude with a discussion and summary.

OVERVIEW OF METHODS AND PRACTICES
There is a variety of methods for obtaining emotional voice data (see, e.g. [4]). Many unanswered questions exist, including order of magnitude estimates for the number of subjects and duration of recordings needed.

Nonetheless, existing work is well-described by three characteristics. First, emotional conditions are used where emotions usually vary in arousal (intensity) and valence (positive or negative), and may be induced, "natural" (i.e. emotional, but not induced in a controlled way), or acted. Second, recorded speech for each condition is obtained in segments of one second to two minutes of recorded data per subject per condition.. Subjects number from about six to hundreds. Third, recordings are labeled in at least three ways: condition stimulus label (for example "happy video and reading"), coding (a human listener's annotations), or self report (via completion of a questionnaire or survey). We now address each of the three issues, noting some of the ambiguities and proposing solutions.

EMOTIONAL CONDITIONS: WHICH, HOW, HOW MUCH?

Which and How Many Emotions?
First, a decision on which (and therefore how many) emotions to focus on? About five emotions are often used, and the most popular include: joy, sadness, anger, neutrality, fear, annoyance, and surprise. We do not make a claim about which emotions should be used, though we do propose that a consistent definition be used, placing emotions on the three dimensional axes of valence, pleasure, and power. We also note that in general positive valence conditions should be presented before negative because the latter tend to persist [Nass, Personal Communication].

Acted, Induced, or "Natural?"

The discussion of the value of induced versus acted vocal emotion (see, e.g. [1]), indicates that induced or "found" emotion (e.g. observed) is superior because it represents how people actually speak. For applications such as interactive media where we must model genuine emotions, we agree that induced or observed emotions should be used. There are, however, situations where acted speech is preferable, such as when the application will ultimately use acted emotional speech for testing data [13]. This would occur, for example, in an automatic indexing application for television programs or films.

There are a number of ways to produce or obtain induced or found emotion data. In [4], the researchers use 10-60 sec clips of people in emotional discussions, either live with members of a research team or as seen on a TV program. The Reading / Leeds Corpus similarly uses emotional data from television programs, not obtained in an experimental context. In [9], emotions were induced by having subjects do readings of emotional passages (7-8 sentences).

Using video clips as stimuli – rather than the data itself – provides another option. Detenber & Reeves [3] suggest that image motion in video clips influences self-report of emotional arousal, but not valence. The study showed that still versions of pictures elicited greater arousal than did the moving versions of the same pictures. Two pilot studies also showed the effects of video clips. The Detenber clips and a fifteen-minute movie clip paired with a frustrating mathematical task were each found to have significant effect on subjects as measured by the 32 term Differential Emotional Scale (DES) in a driving experiment [10]. Similarly, two contrasting seven minute film clips followed by condition-matched vocalized readings and constrained utterances were found to have significant effects on subjects as measured on an abbreviated (20 term) Positive and Negative Affect Schedule (PANAS) in a voice emotion study [8]. The effects were observed both for vocalized readings of given text and for open ended utterances about the given readings or videos.

Depending on the emotion to induce, video clips may not always be optimal. Several researchers have reported difficulty eliciting high levels of reported anger [6]. Moreover, films designed to elicit anger states often turn out to elicit a blend of negative emotions, including related states such as disgust and sadness. Films are at a disadvantage relative to techniques that induce anger through interpersonal situations. The explanation for this is most likely that anger requires a high level of personal engagement and/or immediacy and this is hard to achieve with a film. Other methods to induce anger include computer based tasks, such as the Stroop color-naming and math tasks [14].

To induce most emotions we propose using film clips ranging in length from five to fifteen minutes. We especially propose using film clips that have been rated for valence, arousal and power (e.g. the Detenber film clips). To induce anger and frustration, we propose using computer and visualization tasks in addition to film clips.

How much data and how many subjects?

We are presently investigating the minimum amount of data to collect. In a pilot study [8], a two minute baseline recording improved machine detection of voice emotion. (We are unaware of other speaker-dependent systems that use induced rather than acted emotion.) Specifically, detection of happiness versus sadness improved from 67% to 89% when an individual baseline was included for each subject. The quantity of data and subjects needed is still an open question, and one we hope will be addressed by the framework we propose and by additional data mining were are presently doing on the voice emotion pilot study data.

LABELING OF DATA: EMOTIONAL CONDITION, CODING, AFFECT SCALE, OR BIOMETRICS?

There are at least four ways to measure and label the effect of emotional manipulation. We describe each below, highlighting strengths and weaknesses.

Emotional Condition Labeling

A *condition stimulus label* may simply be assigned to each voice recording. This is suitable for acted and induced emotional conditions. This labeling scheme is used by McGillolway et al [9], in a pilot study in which twenty raters confirmed that the texts were "suited to expressing the relevant emotion;" no other labeling such as human coding or self-report was used. Condition stimulus labels cannot be applied if a study uses "found" emotional data.

"Coding" by Human Annotators

Human *coders* may listen to recordings and label each portion's emotional content. This allows fine-granular time-aligned emotion labeling. It does not take into consideration that individuals manifest voice emotion in different ways. Human coding is often used for non-experimental data, such as film clips. Such coding is non-trivial, and can include continuous graphical representation [2], manual annotation, or hierarchical labeling of single or "blended" Major and Minor emotions [15]. In the latter, 85% intra-coder consistency was measured for some twenty emotional class labels, a strong showing.

Self-report by Subjects

Subjects may *self report*, either by completing some measure like the Positive and Negative Affect Schedule (PANAS) or by explicitly describing the emotions they felt during the voice recoding. A complication of self-report is reliance on memory since subjects report on their emotions by recall after specking. Similarly, subjects cannot be expected to self-report on emotions with the same granularity as human coders.

There are however also advantages to self-reporting. Real rather than masked emotions are self-reported, whereas this deception might not be picked up by a human coder that instead might indicate the masked emotion. Another advantage is that self-reports can be completely automated through a web interface or a paper questionnaire. So, self-reporting has the potential to allow large numbers of subjects to be efficiently recorded in relatively short time.

In a recent study by Master et al [8], the PANAS data showed strong correlation with condition and machine-detected voice emotion. The PANAS scale is based on the prompt, "indicate to what extent you feel this way right now," followed by a list of 20 emotion adjectives. The adjectives are each associated with a five-point Likert scale anchored by "very slightly or not at all" (=1) to "Extremely" (=5). A factor analysis of the 20 adjectives scale revealed one dominant factor (lambda=6.45; R^2=.32). The items in the factor were combined into an additive index, which was reliable (Cronbach's alpha=.65). A paired-comparison, two-tailed t-test demonstrated that the positive manipulation led to much higher levels of positive arousal, M=3.02, SD=0.79, than did the negative manipulation, M=2.08, SD=0.48, $t(17)=4.65$, $p<.001$, suggesting that the manipulation was clearly successful.

Physiological Data

Physiological sensors are often used in media studies to assess emotional responses, and have been used in at least one voice emotion (stressed / non-stressed) study [12]. Skin conductance and heart rate are most often used since these measures have shown to be associated with two primary emotion dimensions. Skin conductance (GSR) measures arousal and when people are aroused sweat glands become active and GSR increases in frequency and amplitude [7]. Heart rate measures valence, and there is an increase in heart rate in response to pleasant stimuli, and a decrease in response to unpleasant stimuli [5]. Physiological data can be used to correlate data from other sources, however, collecting and analyzing the data increases cost and effort substantially.

Recommendation and Comment

We presently advocate the use of the PANAS (or a similar scale), especially due to its efficient and automated implementation. As a final note on this topic, we observe that some research has referred to machine-detected voice features such as fundamental frequency as labeling or indexing data. We believe that features that a machine can only estimate should be separated from those that are considered "known by humans." Otherwise, the problem of machine recognition of emotion becomes circular.

DATABASE IMPLEMENTATION

Online Setup

An online implementation of a database is desirable. All data acquisition described thus far may be implemented via a web interface. Video stimuli, PANAS surveys and voice recording may all be done on a modern computer. Given the use of high quality voice transmission, (e.g. Google Talk), it also seems possible that real-time voice data transmission could be done over fast internet connections. For slower connections, videos for inducement and recordings of subjects could be respectively downloaded and uploaded "offline".

If web interfaces could not be used, we suggest using a standard voice lab. Recording quality should be high and various types of noise (cell phone coding, land line coding, and car noise) can be simulated later. Recording quality of 32 kHz sampling, 16 bit resolution, 50 dB Signal-to-Noise Ratio, should be sufficient for all known automatic voice recognition algorithms. All interfaces that record data should meet this level of quality and automatic tests to ensure this should be built into the web or other interface.

Data Structures

The automated system should construct a metadata spreadsheet file (or spreadsheet row) for each voice recording, which includes the items listed below. This will enable data to be sorted not just by condition, but by length of data or subject emotion self-report.

- Basic data about the recording: subject ID, approximate date, and length of utterance.
- A code or note indicating the type of emotional stimulus used. (Were videos, readings, or both used? Or was the voice data simply stripped from a TV show?)
- "Multiple indexing" for the condition, PANAS / DES data, and (optional) human coding data. This way, the metadata is backwards compatible with experiments that do not use labels other than condition for the data.
- A live "link" to the recording itself.

PRIVACY CONCERNS

An often raised issue in voice emotion research is privacy: how can subjects talk about emotional subjects and have this data released to the public? We make a few observations that could reduce privacy concerns. First, we recall that the NIST Speaker ID evaluations (in which systems attempt to automatically detect speaker identity 11]) use data obtained from consenting subjects engaged in phone conversations. The data is obtained by the Language Data Consortium (LDC) at the University of Pennsylvania. Subjects are aware that the data will be annotated for Speaker ID research, and that the recordings could be played publicly. That many subjects were willing to participate is an indication that privacy might not be a large concern. The general model we propose for data collection and storage is the LDC; we also propose allowing subjects to cancel permission to use their recordings within some time window, say one week, of the recordings being made.

An example study using highly emotional data provides an example for how to reduce privacy concerns. In [15], twenty hours of data was used from an emergency call center. The study followed three principles to ensure privacy: "anonymity of the callers, the privacy of personal information and the non-diffusion of the corpus and annotations." In the proposed context, subjects could be instructed to avoid identifying themselves or others (or locations), using only first names or fake names. To support this, subjects could be allowed to listen to their recordings before they were submitted.

DISCUSSION AND SUMMARY

We have introduced a framework for creating an emotional voice database. The proposed data collection would use video stimuli for most emotional conditions, and tasks for others. We propose data labeling not just by condition, but by subject self-report via a questionnaire taken just after the recording is made. Significantly, all aspects of the proposed data collection could be automated in a web interface. Open questions still remain regarding the number of subjects and length of recordings needed, though we hope the proposed database could be used to help answer them.

ACKNOWLEDGMENTS

We thank the support staff at CHIMe and LREC for their assistance as well as our reviewers for their helpful comments. We also thank Nissan for their collaborative efforts on a Japanese version of the voice emotion pilot study, which will be detailed in a forthcoming publication.

REFERENCES

1. Campbell, N. Databases of Emotional Speech. In *Proc. ISCA Workshop on Speech and Emotion,* QUB, 2000. http://www.qub.ac.uk/en/isca/proceedings/pdfs/campbell.pdf

2. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M. 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time. In ISCA Workshop on Speech & Emotion, QUB, 2000. http://www.qub.ac.uk/en/isca/proceedings/pdfs/cowie_et_al.pdf

3. Detenber, B. H., & Reeves, B. A bio-informational theory of emotion: Motion and image size effects on viewers. *Journal of Communication,* Oxford UP, 1996, 66-84.

4. Douglas-Cowie, E., Cowie, R., and Schröder, M. A new emotion database: considerations, sources and scope. In *ISCA Workshop on Speech & Emotion,* QUB, 2000. http://www.dfki.de/~schroed/articles/douglascowieetal2000.pdf

5. Fitzgibbons, L. & Simons, R. F. Affective response to color-slide stimuli in subjects with physical anhedonia: A three-systems analysis. *Psychophysiology,* 29(6) (1992), 613-620.

6. Gross, J., Levenson, R. Emotion elicitation using films. *Cognition & Emtotion,* 9 (1995), 87 – 108.

7. Hopkins, R. and Fletcher, J. E. Electrodermal measurement: Particularly effective for forecasting message influence on sales appeal. In A. Lang (Ed.), *Measuring psychological responses to media,* Lawrence Erlbaum Associates, 1994, 113-132.

8. Master, A., Deng, P., Richards, K., and Nass, C. (2005). Inducing and Detecting Emotion in Voice. Presented at CIS Poster Session, Stanford University, 2005. http://ccrma.stanford.edu/~asmaster/RA/CISposterslides2.ppt

9. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, C.C.A.M., Westerdijk, M.J.D., and Stroeve, S.H. Approaching automatic recognition of emotion from voice: a rough benchmark. In *Proc. ISCA workshop on Speech and Emotion,* QUB, 2000. http://www.qub.ac.uk/en/isca/proceedings/pdfs/mcgilloway.pdf

10. Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., et al. Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion. In *Proc. CHI 2005,* ACM Press (2005).

11. National Institute of Standards and Technology. *2006 NIST Speaker Recognition Evaluation.* http://www.nist.gov/speech/tests/spk/2006/.

12. Rahurkar, M., Hansen, J. H. L., Meyerhoff, J., Saviolakis, G., and Koenig, M. Frequency Distribution Based Weighted Sub-band Approach for Classification of Emotional/Stressful Content in Speech. In *EUROSPEECH-2003 / INTERSPEECH-2003*, 721-724.

13. Seppänen, T., Toivanen, J., and Väyrynen, E. MediaTeam Speech Corpus: a first large Finnish emotional speech database. *University of Oulu, Finland.* http://www.mediateam.oulu.fi/publications/pdf/394.pd

14. Stafford, T. Integrating psychological and neuroscientific constraints in models of Stroop processing and action selection. PhD thesis, Sheffield University, 2003.

15. Vidrascu, L. and Devillers, L. Real-Life Emotion Representation and Detection in Call Centers Data. In: Tao, J., Tan, T., and Picard, R. W. (Eds.): *ACII 2005*, LNCS 3784 (2005), 739-746.

Affect Sensing using Lexical Means: Comparison of a Corpus with Movie Reviews and a Corpus with Natural Language Dialogues

Alexander Osherenko

Multimedia Concepts and Applications, Faculty of Applied Informatics
University of Augsburg, Germany
osherenko@informatik.uni-augsburg.de

ABSTRACT

Affect sensing using pure lexical means is a problem that has been standing in the centre of attention of research community for a very long time and there exist promising approaches for solving this problem. The questions that have been investigated so far refer to the emotional meaning of a text is and its computation. Although the answers to these questions are certainly important, it is probably too premature to provide them without knowing differences between corpora used for lexical affect sensing. This contribution provides a comparison of two emotional textual corpora – a carefully built and worded movie review corpus and the spontaneous SAL corpus – and presents features for the classification of affective utterances.

Keywords

Lexical affect sensing, emotion research.

INTRODUCTION

Affect sensing is an important field of study. Different efforts have been already made in this research area studying textual, facial, physiological aspects etc. [3]. There is a great application potential of affect sensing, such as storytelling, interface agent applications, learning systems, telephony etc.

This paper compares affect sensing using pure textual means in two text corpora that can be seen in this context as two extremes of affect expressivity – a carefully built and worded movie review corpus and a spontaneous speech corpus. Thus, the question arises of whether the two corpora can be still analyzed using the same features or whether the feature set should vary according to some specific properties of the particular corpus.

This paper doesn't discuss advantages or drawbacks of particular groups of approaches to lexical affect sensing (keyword spotting, lexical affinity, statistical natural language processing, hand-crafted approach). For exact description of these topics see [4].

CORPORA

Two textual corpora are compared – a corpus with movie reviews [1] and the SAL corpus [6].

The movie review corpus contains 84 film reviews rated by a human and mapped onto classes from one to three stars. There are 28 reviews for each rating in the corpus (Fig. 1).

> [Rating: 3]
>
> Simply put, Sofia Coppola's Lost in Translation is an amazing motion picture. There may be some controversy over whether she truly wrote the screenplay on her own (there are sequences that argue that she at least had help from someone with a little more experience in life and marriage), but that doesn't impact the final analysis. …
> … As good as The Virgin Suicides is, Lost in Translation is superior in almost every way. When Top 10 lists are released at the end of the year, this title will feature prominently on a number of them (including mine).

Figure 1: An example of a movie review

The SAL corpus (Sensitive Artificial Listener) is a set of affective dialogues presenting four characters equipped with different responses in differing emotional states: optimistic and outgoing (Poppy), confrontational and argumentative (Spike), pragmatic and practical (Prudence), depressing and gloomy (Obadiah). These dialogues are annotated by four critics (cc, dr, em, jd) with FEELTRACE data [2]. The corpus consists of 27 dialogues (569 annotated utterances from critic cc). The FEELTRACE data for the evaluation affect score are mapped onto three categories (“negative”, “neutral”, “positive”) to facilitate a comparison of the two corpora.

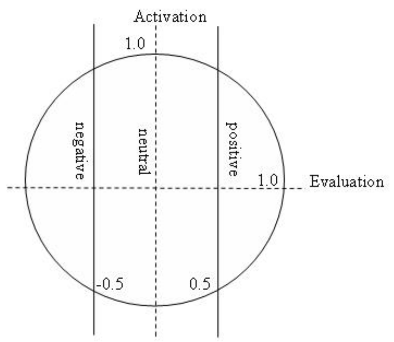


Figure 2: Affect segments representing evaluation

The FEELTRACE data are mapped onto affect segments as shown in Figure 2 (“positive”, “neutral”, and “negative” according to the value of the evaluation dimension). Radius 1 is predetermined by FEELTRACE, whereas the values -0.5 and 0.5 are chosen for the sake of symmetry. Note that there are also such utterances that change their affect meaning from positive to negative or vice versa. These inconsistent utterances occur very seldom and can be excluded from the analysis without significantly influencing the results.
The distribution of affect segments as annotated by critic cc is the following: 'positive': 62, 'neutral': 447, 'negative': 59, 'undefined': 1 where the pairs show an affect segment and the number of utterances of this segment in dialogues. Affect segment “undefined” denotes inconsistent affect segments. Sample utterances from the SAL corpus are shown in Fig. 3.

[Affect segment: positive]
– Oh I'm pretty good I guess. It's nice to hear a cheery voice though.
[Affect segment: neutral]
– Well, there are some real idiots on the road aren't there.
[Affect segment: negative]
– Well that's true too, but if you dwell on that your not gonna get by life in a very (laugh) positive frame of mind.

Figure 3: Examples of SAL utterances

Note that the examples above are provided without a context in order to stress its importance in emotional comprehension – in some cases it is clear what meaning an utterance has, in some cases yet not clear.
Differences between the movie review corpus and the SAL corpus can be expressed as follows:

1. The movie reviews corpus contains affect words that can convey emotional meaning whereas the SAL corpus doesn’t normally have such obvious signs of emotional meaning.
2. The length of emotional texts in the movie reviews corpus is rather long (about 1000 signs) whereas the utterances of the SAL corpus can contain only one word.
3. The movie reviews are grammatically correct sentences whereas the utterances in the SAL corpus can be grammatically incorrect, contain repairs, repetitions and inexact wordings.
4. The movie reviews texts are usually consistent regarding the expression of reviewer’s opinion about the movie whereas some utterances in the SAL corpus can be inconsistent (can convey contradicting emotional meaning).
5. There is no connection between particular reviews in the movie review corpus whereas affect segment values for utterances in a dialogue can be seen as a continuous stream of information between utterances that influences the set of possible features e.g. the feature set can contain a feature for the previous average value.
6. A movie review is composed and rated by the same person whereas test persons and critics in the SAL corpus are different people. This raises a lot of questions e.g. if a critic has the same or at least a comparable impression of verbal entities regarding their affect meaning as a person he/she rates.

Due to these differences, we also expected different results for automated methods to affect sensing.

FEATURE EXTRACTION

For each corpus, we computed the following features:

1. Affect word features. Affect words from the Whissell’s dictionary of affect as in the lexical affinity approach. Chosen affect words can belong to every part-of-speech (POS) group out of 40 (see below), be it an adverb, a pronoun whatsoever. The set of affect words doesn’t contain every affect word that can appear in an affect utterance in every grammatical form. Section 4 shows that it is fairly sufficient to have only particular words in the learning model (cf. Table 2).
2. Average value features. Average values of affect scores of affect words found in a particular utterance.
3. Previous average (only in the SAL corpus). Average values of affect scores from previous utterance in a dialogue.
4. Part-of-speech (POS) tags and their combinations [5]. The features describe the number of corresponding POS groups in an utterance.

CLASSIFICATION

The extracted features are evaluated using the WEKA machine learning toolkit using the SMO classifier – a WEKA analogue of SVM [8].

Feature calculation

1. Affect word features. We performed tests with a different number of affect word features. In particular, we used all affect word features (8574) or the affect word features with the highest evaluation scores (positive or negative) according to four thresholds resulting into sets of 2451, 1307, 950 or 765 features.
2. Average value features. Average values for evaluation, activation, and imagery are calculated from the emotional scores of affect words in the current affect text.
3. Previous average – the average from the previous affect

text (only in the SAL corpus).

4. Part-of-speech (POS) tags and their combinations. The number of the features in this group is 36 conventional + 4 derived combinations (“NN or NNS or NNP or NNPS”; “VB or VBD or VBG or VBN or VBP or VBZ”; “JJ or JJR or JJS”; “RB or RBR or RBS”) and 1600 possible combinations. The values of these features are the numbers of POS or combination occurrences in an affect text (movie review or an utterance).

Results

Table 1 shows results of affect sensing in the movie review corpus and the SAL corpus. The values are averaged over affect classes.

s#	Movie Features	R (%)	P (%)	SAL Features	R (%)	P (%)
1	#W: 8574	54.76	54.54	#W: 8574	43.74	46.43
2	#W: 2451	54.76	56.66	#W: 2451	40.30	51.01
3	#W: 1307	59.52	60.95	#W: 1307	36.52	42.87
4	#W: 950	59.52	61.83	#W: 950	39.33	53.29
5	#W: 765	54.76	55.95	#W: 765	38.61	54.59
6	#W: 8574 #POS: 40	54.76	54.54	#W: 8574 #POS: 40	42.73	44.42
7	#W: 2451 #POS: 40	55.95	56.92	#W: 2451 #POS: 40	39.98	47.90
8	#W: 1307 #POS: 40	63.09	62.96	#W: 1307 #POS: 40	37.62	47.58
9	#W: 950 #POS: 40	59.52	61.82	#W: 950 #POS: 40	38.31	49.70
10	#W: 765 #POS: 40	55.95	56.38	#W: 765 #POS: 40	40.25	58.56
11	#W: 8574 #POS: 1640	59.52	57.91	#W: 8574 #POS: 1640	41.48	45.36
12	#W: 2451 #POS: 1640	63.09	62.82	#W: 2451 #POS:1640	40.57	46.82
13	#W: 1307 #POS: 1640	64.28	64.45	#W: 1307 #POS:1640	40.21	45.55
14	#W: 950 #POS: 1640	63.09	63.42	#W: 950 #POS:1640	38.70	43.76
15	#W: 765 #POS: 1640	64.28	65.0	#W: 765 #POS: 1640	36.12	38.63
16	#W: 8574 A: 3	53.57	53.61	#W: 8574 A: 3	42.57	44.76
17	#W: 2451 A: 3	53.57	55.14	#W: 2451 A: 3	39.89	50.07
18	#W: 1307 A: 3	57.14	58.89	#W: 1307 A: 3	36.67	44.23
19	#W: 950 A: 3	60.71	63.51	#W: 950 A: 3	38.84	53.39
20	#W: 765 A: 3	50.0	52.36	#W: 765 A: 3	38.69	55.88
21				#W: 8574 A: 3 Pr: 3	42.30	44.86
22				#W: 2451 A: 3 Pr: 3	40.38	51.41
23				#W: 1307 A: 3 Pr: 3	36.67	43.92
24				#W: 950 A: 3 Pr: 3	38.72	50.20
25				#W: 765 A: 3 Pr: 3	38.76	56.90

Table 1: Results of affect sensing

The Features column designates the features used where “#” column shows the row number, #W is the number of affect words from the Whissell’s dictionary of affect, A represents the average score feature and Pr the previous average score (only in the SAL corpus). The R, P columns are the conventional recall and precision measures. The number after the colon represents the number of values for this feature.

Feature comparison

It is impossible to give a clear answer to the question of whether affect words influence affect sensing in an obvious manner (cf. Table 1). Even with a smaller intersection of words from the Whissell’s dictionary of affect it is possible to achieve better results for affect sensing (cf. rows 2 and 3; rows 3 and 4; rows 7 and 8 in the Movie corpus or rows 9 and 10 in the SAL corpus). Note that the result values don’t introduce monotonicity regarding the number of affect words (higher words number means higher precision and recall values) e.g. in the SAL corpus row 1 contains the highest value of the recall measure whereas row 5 contains the highest value of the precision measure.

Table 2 presents numerical characteristics of occurrences of affect words in the Whissell’s dictionary of affect. Note the lowest intersection of the set of the words from the Whissell’s dictionary of affect and the set of words from the movie review corpus is 5.08% whereas the intersection of the set of the words from the Whissell’s dictionary of affect and the set of words from the SAL utterances is not greater than 2.04%.

#W	#M.I.	#M.A.	M./W. (%)	#S.I.	#S.A.	S./W. (%)
8574	4815	6647	56.15	175	25	2.04
2451	1395	10067	16.27	41	159	0.47
1307	750	10712	8.74	21	179	0.24
950	541	10921	6.30	17	183	0.19
765	436	11026	5.08	10	190	0.11

Table 2: Intersections of affect words

where #W is the number of affect words from the Whissell's dictionary of affect, #M.I. – the intersection of the set of words from the movie review corpus and the set from the Whissell's dictionary of affect, #M.A. – the number of words from the movie review corpus that are absent in the set of affect words from the Whissell's dictionary of affect, M./W. – the ratio of the set of words from the movie review corpus and the dictionary of affect, #S.I. – the intersection of the set of words from the SAL scenario and the set from the Whissell's dictionary of affect, #S.A. is the number of words from the SAL that are absent in the set of affect words from the Whissell's dictionary of affect, S./W. is the ratio of the set of words from the SAL corpus and the dictionary of affect.

Our classification algorithm provided better results for the movie review corpus (the precision measure for a three class problem – "negative", "neutral", "positive" – is on average 58.98%) than for the SAL corpus (the precision measure for a three class problem – "*", "**", "***"– is on average 48.20%) independently of the number of affect word features. Another interesting observation is that a reduction of affect word features for the movie corpus yielded better results. Only when reducing the set to 765 affect words the results were similar again to the results obtained for 8574 words. In contrast, the best results for the SAL corpus were obtained when using the full set of 8574 features.

The POS feature improves affect sensing in the movie review corpus and at least influences the processing in the SAL corpus (cf. rows 1-5 with rows 11-15) although the influence of POS groups on affect sensing in the SAL corpus is less obvious. A higher number of POS groups (1640 vs. 40) always led to higher recall and precision values.

The average affect score features had no essential influence on the recognition rate. In the SAL corpus, for example, using 1307 affect word features the recall measure was 36.52% and precision measure 42.87%, whereas using 1307 affect word features and the average score features – 36.67% and 44.23% resp. (cf. rows 1-5 with rows 16-20 in both corpora). Note the results in row 19 (38.84%-53.39%).

The previous average feature influences affect sensing. Worse results are achieved in rows 18 and 23, better results in rows 20 and 25.

In sum, all the features above can positively influence classification results although it is sometimes intuitively difficult to find a correct balance between feature values.

CONCLUSION

This contribution conducted a comparison of two emotional textual corpora. It also provided features for classifying affect utterances in these corpora. The best classification results for three affect segments are – for the recall measure 64.28% and 43.74%, for the precision value 65% and 55.88% in the movie review corpus and SAL corpus respectively.

In our future research, the following areas will be investigated more thoroughly taking into consideration the differences between emotional text corpora:

1. Using stemming algorithms to recognize affect word features
2. Testing of different algorithms for feature selection
3. Considering a longer dialogue history both for affects score features as well as POS features.

REFERENCES

1. Berardinelli, J. (2006). URL: http://movie-reviews.colossus.net.
2. Cowie, R. et al. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In: Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland. pp. 19–24.
3. HUMAINE (2006). Research on Emotions and Human-Machine Interaction. URL: emotion-research.net.
4. Liu, H. & Lieberman, H. & Selker, T. A (2003). Model of Textual Affect Sensing using Real-World Knowledge. Proceedings of IUI 2003. Miami, Florida.
5. Mitchell P. M. & Santorini B. & Marcinkiewicz M.A.. (1993). Building a large annotated corpus of english: The penn treebank. Computational Linguistics, 19:313-330.
6. Semantic Affect Listener – SAL (2006). URL: http://emotion-research.net/ws/wp5/ellsal.ppt.
7. Whissell C. (1989). The dictionary of Affect in Language. In: R Plutchik & H Kellerman ed. Emotion: Theory, research and experience vol. 4 Academic Press: New York.
8. Witten I. H. & Frank E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition.

An Italian Database of Emotional Speech and Facial Expressions

N. Mana[*], P. Cosi[†], G. Tisato[†], F. Cavicchio[†], E. Caldognetto Magno[†], F. Pianesi[*]

[*]ITC-irst

Center for Scientific and Technological Research
via Sommarive, 18 – 38050 Povo (Trento), Italy
{mana, pianesi}@itc.it

[†]ISTC-CNR
Institute of Cognitive Sciences and Technology
via Anghinoni, 10 – 35121 Padova, Italy
{cosi, tisato, cavicchio, magno}@pd.istc.cnr.it

ABSTRACT

This paper presents an Italian database of acted emotional speech and facial expressions. New data regarding the transition between emotional states has been collected. Although acted expressions have intrinsic limitations related to their naturalness, this method can be convenient for speech and faces synthesis and within evaluation frameworks. Using motion capture is a good method to get precise information on data for playing back them on facial model and also to build specific animation engine. The procedure to adapt the recorded data to a MPEG-4 compliant facial animation model will be described.

Keywords

Emotional speech and facial animation, 3D motion capture.

INTRODUCTION

During the last years there has been a growing interest in systems working on emotions. Many works addressed recognition and synthesis of emotions through vocal and facial signs. In this perspective, emotion corpora become fundamental to perform conceptual analyses, to develop emotion recognition and synthesis systems (especially in case of data-driven systems) both for speech and face, and to test emotion-oriented tools and applications [3].

In acquiring basic data, four main types of source are commonly used: a) spontaneous emotions; b) inducted emotions; c) acted emotions; and, d) application-driven emotions [1].

In this paper we present an audio/video database of acted emotional speech and facial expressions.

DATA COLLECTION

For the data collection a professional actor (male, 25 years old) was employed. He was instructed to utter short non-sense words with various emotional expressions and intensities.

Collected Data

The first part of the database includes "Isolated Emotions", i.e. a set of Italian non-sense words, acted with different emotions. These words, representing Vocal – Consonant – Vocal (VCV) sequences (specifically /aba/, /ada/, /aLA/, /adZa/, /ala/, /ana/, /ava/), cover the seven basic viseme[1] classes for Italian [7].

Each VCV sequence was acted with six emotional states, corresponding to the Ekman's set [5] – Anger, Disgust, Fear, Happiness, Sadness, and Surprise – plus the additional 'Neutral' state. Each emotion was acted with three different intensity levels (Low, Medium, High). Examples of the six emotional and neutral states are shown in Figure 1.

"anger" "disgust" "fear" "happiness" "neutral" "sadness" "surprise"

Figure 1. Examples of emotional facial expressions during speech.

The second part of the database includes "Combined Emotions", i.e. VCV-VCV sequences. In this case, in order to get significant examples of transitions from an emotional state to another during speech, the non-sense words were acted in pairs, each one with a different emotional state (Neutral, Anger, Happiness, Surprise), at medium intensity.

Finally, the third part includes examples of a long sentence with a good coverage of Italian phonemes ("Il fabbro lavora con forza usando il martello e la tenaglia"; lit. "the smith works with strength using the hammer and

[1] A viseme is a "visual phoneme", i.e. the visual equivalent of a phoneme (unit of sound) in spoken language. Phonemes can be clustered according to their visual similarity. So, e.g. /p/ and /b/ are phonetically different but they belong to the same viseme class given that, from a visual viewpoint, cannot be distinguished.

the tongs”), acted with the neutral and the six emotional states and three intensities.

Globally 1,657 examples of emotional expressions during the speech have been collected.

Recording: Technical set-up and Procedure

The dynamics of every facial expression during the speech, as well as the respective articulatory and acoustic data, was captured with high precision by means of ELITE [4]. ELITE is an automatic opto-tracking movement analyser for 3D kinematics data acquisition, which also allows synchronous recording of the acoustic signal.

The system records the acoustic signal and tracks the infrared light reflected by 28 small (only 2 mm diameter) passive markers, glued on the actor’s face in different positions. For each marker, displacement and transition speed from a position to another, are recorded every 1/10th on a second by two cameras. In this way the dynamics of every facial expression is captured with high accuracy (100 Hz sampling rate, maximal error of 0.1 mm for a 28x28x28 cm cube).

A speaker announced to the actor the short word to be played, followed by the corresponding emotional state and intensity, according to the following scheme:

< short word><emotional state><intensity>

For example: *Ava, Disgust, High.*

To ensure an easier detection of the starting and ending point of the emotional expression, the actor uttered two additional Italian words, “chiudo” (lit. I close) and “punto” (point), respectively before and after the short word played emotionally. For example: “*chiudo*” [*ava*]$_{\text{Disgust, High}}$ “*punto*”

The order of the sequences to act was generated randomly, so that the actor could not be conditioned. Only the intensity sequence (Low, Medium and High) was kept fixed so that the actor could use the first intensity as reference and generate the second and the third expression increasing the intensity progressively.

DATA PROCESSING

Given the recordings of the marker movements for each emotional expression, two kinds of data processing were accomplished: a) a 3D reconstruction of the marker trajectories; and b) a conversion of these trajectories into Facial Animation Parameters.

3D reconstruction

Starting from the recordings captured by the two cameras of the ELITE system, it was necessary to build the 3D trajectories corresponding to the articulatory movements of the 28 markers on the face. For this task a MatLab© software (called “Track” [10]), specifically developed by ISTC-CNR, has been used.

Track allows to construct the 3D trajectories of markers, frame by frame, starting from the 2D recordings. For doing that, it needs a “reference model”, i.e. a file in which a name is associated to each marker, so that an unambiguous linking between marker and its position is created for each camera (TV1 and TV2), as showed in Figure 2.

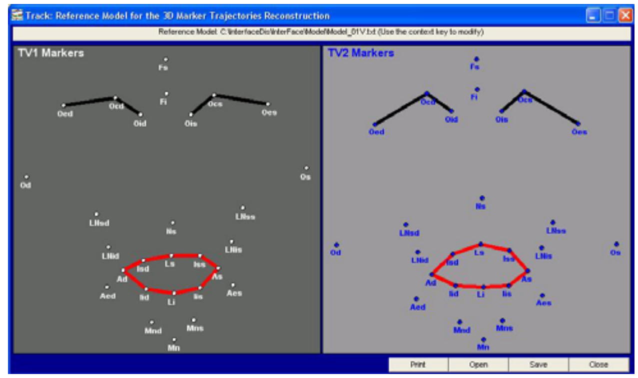


Figure 2. Definition of the “Reference Model” for the 3D reconstruction in the Track system

The system output is a set of values corresponding to the x,y,z coordinates for each marker, frame by frame.

The marker identification and the reference space deformation problem have been exceeded with an algorithm based on the Singular Value Decomposition (SVD) which has the intrinsic advantage to operate an error minimization while calculating the roto-translation, even independently from using a perfect undeformable reference space. The produced FAP-stream takes into account the roto-translation and the scale factors of the head that has to be animated, thus allowing a correct data-driven synthesis of whichever MPEG-4 compatible agent.

Conversion to FAPs

Given the marker trajectories in the 3D space (x,y,z coordinates), these were converted into FAPs (Facial Animation Parameters), according to MPEG4 standard[2]. FAPs are based on the study of minimal perceptible actions and are closely related to muscle actions involved in a facial expression. MPEG4 animation uses a pseudo-muscle approach, in which the muscle contractions are obtained through the deformation of the polygonal network around feature points. Each feature point corresponds to skin muscle attachment. The deformation is performed in a zone of influence that has an ellipsoid shape whose centroid is a feature point. The displacement of points within this area of influence obeys to a deformation function that is function of the distance between the points and the feature point. MPEG-4 defines 84 Feature Points (FPs) that describe the shape of a

2 MPEG4 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) [4], widely used in 3D animation. More details may be found at MPEG4 web page: Hhttp://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm

standard face and should be defined for every face model. These points are used for defining animation parameters as well as calibrating the models when exchanged between different players. Two sets of parameters describe and animate the 3D facial model: facial animation parameter set (FAPs) and facial definition parameters (FDPs). FAPs define the facial actions, while FDPs define the shape of the model.

FAPs have to be calibrated prior to use them on a specific face model. For this reason, FAPs are expressed in normalized units called FAPUs (Facial Animation Parameter Units) which are defined as fractions of distances between key facial features (e.g. eye-nose separation). Only FAPUs are specific to the actual 3D face model that is used, while FAPs are independent. That means they can drive different face models, regardless of geometry. As a result, by coding a face model using FPs and FAPUs, developers can freely exchange face models without the problem of calibration and parameterization for animation, and FAPs can be used for different 3D facial models.

When the model has been characterized with its FDPs (namely the model shape has been defined), the animation is obtained by specifying the FAP-stream, i.e. the values of FAPs frame by frame.

The 68 FAPs values, specified in the FAP-stream frame by frame, cause the facial animation by defining the deformation between two frames of animation. Of these 68 values, the first 2 are high level parameters representing visemes and emotions. The remaining 66 are low level FAPs, dealing with specific regions on the face (e.g. bottom of chin, left corner lip, right corner of left eyebrow, etc.). Most of FAPs correspond to an FP, and define translation or rotation on that FP along an axis in three dimensions, while some of the FAPs represent rotation of the head and eyes.

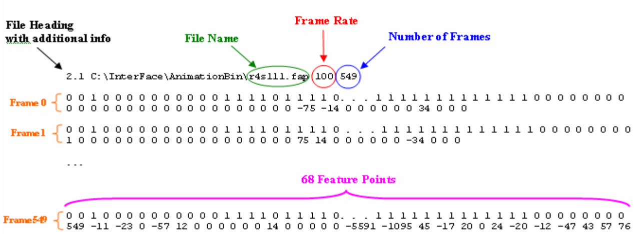


Figure 3. A FAP stream

As shown in Figure 3, in a FAP-stream the relevant information for animation are distributed on two lines: the first line indicates which point is specifically activated in that moment (activation or not is expressed by 0 and 1), while the second one contains the target values, in terms of differences from the previous frame target values. When a FAP is activated (i.e. when its value is not null), the feature point on which the FAP acts is moved in the direction indicated by the FAP itself (up, down, left, right, etc).

By converting the 3D marker trajectories into FAP streams, it is possible to animate the emotional facial expressions played by the actor on any MPEG4-compliant synthetic face, as depicted in Figure 4.

Figure 4. Basic emotions on a synthetic face

DATA ANALYSIS

Emotional Facial Expressions

In order to represent graphically the dynamics of a facial expression, i.e. the marker trajectories frame by frame, we converted the x,y,z coordinate values for each marker in a unique value. For this purpose we used the 3D vector module:

$$|V| = \sqrt{(x^2 + y^2 + z^2)}$$

where x,y,z represent the distance on the three Cartesian axis of a marker with respect to the axis origin.

We can see, for example, the dynamics[3] of "Neutral" in Figure 5 and that one of "Surprise" acted with medium intensity by the actor when uttering "aba" in Figure 6.

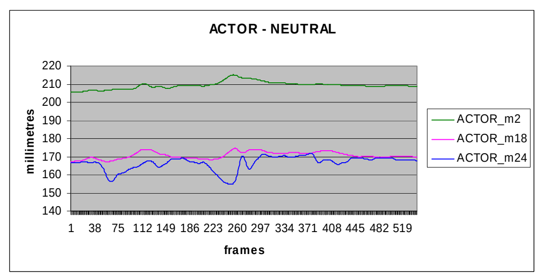


Figure 5. Dynamics of a "Neutral" facial expression

By using this graphical representation, the "prototypical behaviour" of each emotional expression becomes evident. Note, for example, in Figure that the m2 curve presents a pick in correspondence to the maximum value of the surprise expression, as typically it happens[4]. By graphs it is also possible to analyze how emotions affect the speech production. This is particular evident by comparing the dynamics of emotional expressions for a

---

[3] Note that for the sake of graph readability, the dynamics of not all of the 28 markers is showed. The analysis is focused only on marker 2 (m2), marker 18 (m18) and marker 24 (m24), corresponding respectively to "Left Central Eyebrow", "Left Lip Corner" and "Middle Lower Lip" because in some way they are the most significant in capturing facial expression changes (see [8]).

[4] As well known, one of the features characterizing the facial expression of "surprise" is "risen eyebrows"

specific viseme with the dynamics of the same viseme uttered without emotion (neutral state).

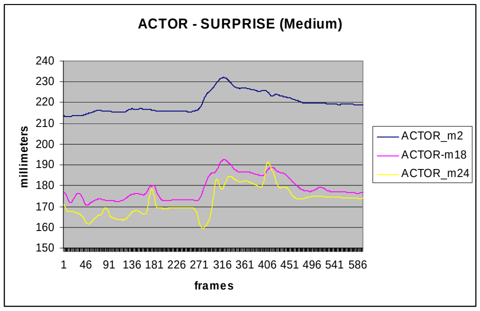


Figure 6. Dynamics of a "Surprise" facial expression

Figure 7 shows, for example, the dynamics of marker 2 for "aba" viseme in neutral and surprise condition[5]. As it is evident by the graph, the curves are strongly affected by the emotion (see the central peak). Furthermore, we can see that the emotional expression tends to be longer than the neutral one.

Emotional Audio/Visual Speech

All these data have been used for building LUCIA [2], an emotional audio/visual talking head that uses 3D polygon models, which are parametrically articulated and deformed by a data/driven-based animation engine [1], and speaks with the Italian version of the FESTIVAL diphone TTS synthesizer [9], appropriately modified with emotive and expressive capabilities.

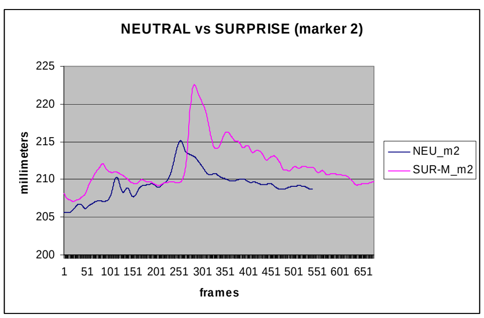


Figure 7. Dynamics of marker-2: neutral vs surprise

CONCLUSIONS

In this paper we have presented an audio/video database of acted emotional speech and facial expressions. Although acted expressions have intrinsic limitations related to their naturalness with respect to spontaneous ones, they can be convenient and suited for specific tasks, such as speech and faces synthesis and within evaluation frameworks.

[5] More examples of several emotional expressions will be showed during the presentation.

ACKNOWLEDGMENTS

The data collection was carried out within the project PF-STAR (**P**reparing **F**uture multi**S**ensorial in**T**erAction **R**esearch), supported by the EU grant IST-2001-37599 (for more details see http://pfstar.itc.it). The authors would like to thank all the people involved in the data collection, and in particular the actor, Fabio Fusco, for his patient cooperation.

REFERENCES

1. Cosi P., Fusaro A., Grigoletto D., and Tisato G., Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes. In *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems"*, Germany (2004).

2. Cosi P., Fusaro A., Tisato G., LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland, September 1–4, 2003, Vol. III, 2269-2272.

3. Cowie R., Douglas-Cowie E. and Cox C., Beyond emotion archetypes: Databases for emotion modelling using neural networks. In *Neural Networks* - Special Issue Emotion and Brain: Understanding Emotions and Modelling their Recognition, 18 (4), (2005), 371-388.

4. Doenges P., Lavagetto F., Ostermann J., Pandzic I.S., and Petajan E., MPEG-4: Audio/Video and Synthetic Graphics/Audio for Mixed Media. In *Image Communications Journal*, 5(4), (1997).

5. Ekman P, An Argument for Basic Emotions. In N.L. Stein, and K. Oatley, editors, *Basic Emotions*, (1992), 169-200.

6. Ferrigno G., and Pedotti A.m ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing. In *IEEE Transactions on Biomedical Engineering, BME-32*, (1985), 943-950.

7. Magno Caldognetto E., Zmarich C., Cosi P. and Ferrero F., Italian Consonantal Visemes: Relationships Between Spatial/temporal Articulatory Characteristics and Coproduced Acoustic Signal. In *Proceedings of AVSP-97*, Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational & Cognitive Science Approaches, Rhodes, Greece (1997).

8. Mana N. Modeling Dynamics of Emotional Facial Expressions in Talking Heads. PhD Dissertation, Trento (Italy), March 2006.

9. Tesser F., Cosi P., Drioli C., Tisato G., Emotional Festival-Mbrola TTS Synthesis. In CD *Proceedings INTERSPEECH 2005*, Lisbon, Portugal, 2005, 505-508.

10. Tisato G., Cosi P., Drioli C., Tesser F., INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads. In CD Proceedings INTERSPEECH 2005, Lisbon, Portugal, 2005, 781-784.

UT-SCOPE – A corpus for Speech under Cognitive/Physical task Stress and Emotion

Vaishnevi S. Varadarajan, John H.L. Hansen, Ikeno Ayako

CRSS: Center for Robust Speech Systems
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688 U.S.A.
Vaishnevi.Varadarajan@colorado.edu, John.Hansen@utdallas.edu, Ikeno.Ayako@utdallas.edu

ABSTRACT[1]

The widespread deployment of automatic speech processing systems over a range of environmental conditions such as in cars, factories, wireless cell/PDA devices, public speeches/lecture halls, and busy office conditions has motivated the need for reliable performance under varying noise and cognitive task task/emotional stress conditions. In this paper, we describe a speech corpus that is focused on speech under stress – cognitive and physical task stress, Lombard effect and emotion. It is suggested that this speech database will complement the SUSAS corpus collected previously, and be useful in establishing a more scientific understanding and algorithm development in the detection, analysis and compensation of speech under stress, emotion and noise from Lombard Effect. This paper will report the initial development of the UT-Scope corpus, and include results on system evaluation.

INTRODUCTION

Maintaining reliable speech system performance for changing environmental and speaker conditions such as background noise, speaker based Lombard effect, speaker stress and emotion is challenging for researchers in speech processing. For example, consider the general scenario illustrated in Fig. 1 which shows speech submitted to a general speech system, which could be for speech recognition, speech coding, speaker recognition, language ID, dialog system, etc. The figure shows a sample of the diversity of factors which impact the acoustic signal prior to being submitted to the speech system. A system trained on neutral speech, will not be able to perform reliably when environmental and speaker specific variations are introduced into the signal. Interactive speech systems are used in a variety of environments that include: mobile phones, PDAs, dialog systems, information access such as Kiosks, command and control or information access for in-vehicle systems. In-Vehicle environments are emerging as one of the primary domains for more effective interactive systems which allow for driver control in entertainment and information, yet maintain a manageable cognitive task load which minimizes distraction for the driver. Recent workshops and published textbooks have addressed the range of issues for more effective interactive systems for in-vehicle applications[7,8,9]. These systems must deal with speech affected by human factors such as physical and cognitive task stress, emotion and Lombard effect due to noise. Cognitive and physical task stress may be due to strenuous driving conditions, sleep deprivation, etc. Emotional conditions such as fear, anxiety, and anger are also common in a number of application environments including call centers, interactive dialogs, aircraft cockpit communications, military environments, 911 emergency calls. It has been shown that recognition rates for speech and speaker ID systems degrade rapidly when a system trained with neutral speech is tested with speech under emotion, stress or Lombard effect data [1]. Table.1 shows results reported in [1] using the SUSAS corpus [2,6]

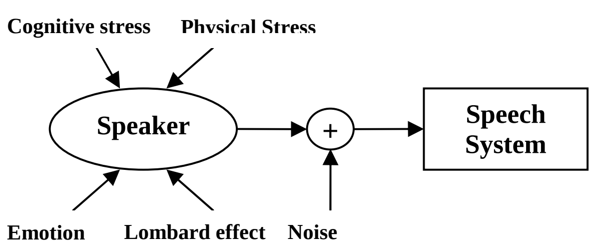


Fig. 1: Generalized Speech System Scenario

Mismatched training refers to training the system with neutral speech and testing with stressed speech. Matched training implies training and testing with stressed speech. From Table 1, it is clear that in addition to noise, human factors such as stress and emotion also effect speech systems.

[1] This work was supported by a Grant from U.S. Air Force under contract No. FA8750-05-C-0029, RADC under contract No. FA8750-05-C-0029, and by the Univ. of Texas at Dallas under Project EMMITT.

In order to deal with the effects of human factors on speech system performance, a detailed analysis of the influence of emotion and stress on human speech is essential. Previous studies in this regard [2,10,3] have been made using corpora containing individual utterances from limited number of speakers. A detailed corpus for speech under noise, emotion and stress, with read and spontaneous sentences from many subjects, with multiple recording sessions, would be useful for further research in this area. This is the motivation for UT-SCOPE: Speech under COgnitive and Physical task stress and Emotion, developed by the Center for Robust Speech Systems, Univ. of Texas at Dallas. In this study, we describe the details of UT-SCOPE and report some preliminary results using this corpus.

Stress/Emotion	Mismatched Training	Matched Training
Neutral	96	96
Angry	34	75
Fast	91	90
Lombard	48	99
Loud	22	81
Slow	90	98
Soft	73	89

Table 1: Recognition accuracy (%) of speaker ID system for matched and mismatched training for speech under emotion and stress

THE PROPOSED DATABASE

UT-SCOPE will consist of read and spontaneous speech from a total of approximately 50-100 speakers. As of April 2006, 35 speakers have been enrolled. The read speech consists of twenty phonetically balanced sentences from the TIMIT database and digit sequences. The speech is obtained under the following conditions:

- Lombard effect due to various noise types at different levels (produced in an ASHA Certified Sound Booth) [35 subjects completed]
- Cognitive stress [in progress]
- Physical task stress [in progress]
- Emotion [in progress]

The acoustic signal is collected using multiple microphones/sensors including: a Physiological sensor (e.g., a throat microphone[P-mic]), a Shure Beta-54 close-talking mic, and a Shure directional far-field microphone using a multi-channel digital audio recorder (Fostex) with channel dependent pre-amp level control. Recent research has shown improvement in automatic stress detection performance using a P-mic with extended algorithm development [4]. In addition to the sensors above, video recordings and biometrics (Heart rate and BP) will be included for physical and cognitive stress conditions, as possible indicators of stress. The following subsections detail the data collection under each of the aforementioned conditions, and the state of the corpus as of April 2006.

Lombard effect

Lombard effect is defined as speech produced by an increased effort to communicate effectively in noisy environments. It has been shown that Lombard effect speech varies from neutral in terms of pitch, intensity, duration, spectral slope, formant location and bandwidth structure, etc.[2,10]. These differences cause a breakdown of speech system performance when systems are trained with neutral but tested with Lombard effect speech. In order to compensate for the variations in Lombard effect speech, it would be interesting to investigate the variations of Lombard effect under different noise types and noise levels.

Fig.2. Top figure shows the three microphones used in performing the data collection. Shure Beta mic., far-field desktop microphone and the throat mic. (from left to right). Bottom picture shows the DAT recorder used in the recordings.

Previous students using SUSAS[6] have considered only a single noise type and level. Thus, a prior knowledge of the noise type and SNR would allow for more advanced and appropriate degree of compensation [10, 11]. In this database, speech under three noise types at three levels is recorded. Pink noise (PNK), large crowd noise (LCR) and noise in a car traveling at 65 mph on a highway (HWY) with windows half open [5] are presented binaurally at different levels using open-air headphones worn by the speaker (i.e., we provide a direct acoustic path for the subject speaking under Lombard effect, yet record a noise-free speech data sequence). The open-air headphones allow the speaker to hear his own voice when speaking. A pure-tone hearing test is conducted on the speaker prior to data acquisition to screen for any potential hearing problems for the subjects under test. The recordings are performed in an acoustically clean double-wall sound booth and data from 35 speakers aged 19-30 are collected. Noise presentation levels varied from 65 dB-SPL to 90 dB-SPL (i.e., 3 levels for each of the 3 noise types, resulting in 9 Lombard conditions).

The speech under Lombard effect consists of 100 read sentences chosen from the TIMIT corpus under neutral condition. 20 phonetically balanced sentences, forming a subset of the aforementioned 100 sentences are used under each of the 9 Lombard effect conditions. The read speech also contains 5 tokens each of the 10 digits (0-9). These text materials were presented using a flat LCD display, with sentences presented in random order for every condition. Additionally, spontaneous speech of one minute duration is recorded by having the subject describe the events/activities/content of pictures and cartoons presented as part of the prompts. As of April 2006, 35 subjects have been collected, each producing 45 min. of data. Of the 35, 20 subjects have multiple recording sessions. We expect to complete 50 subjects by May 2006.

Cognitive stress:

Very often, one might access an automatic speech system while performing a cognitively demanding task like driving a car under heavy traffic conditions. Thus, studying the effect of cognitive stress on speech is of practical significance. In the UT-Scope corpus, speakers will be driving a car-driving simulator using a Sony PlayStation2 in scenarios that require extensive concentration (e.g., the driving simulator has extreme cognitive task conditions). A standard size automobile steering wheel and gas/brake pedal set are used to perform the driving task. A driving seat which incorporates movements (e.g., speed, etc.) from the video player is used. The vibration effect which is transferred to the steering wheel adds to the reality of the simulator. As of April 2006, preliminary data collection from 4 subjects has been conducted to verify the cognitive task load, ease of acquiring task guidelines for the driving simulator, and hardware placement/setup.

Physical task stress:

Physical stress includes factors such as G-force experienced in aircraft cockpits, stress experienced due to high speeds in racing cars etc. In this corpus, speech will be collected while a person operates a stair stepper. Video and biometrics such as heart rate and blood pressure are also recorded for cognitive stress as well as physical task conditions. Speech consisting of spontaneous and read sentences from 75 speakers, with comparable mix of genders will be collected. As of April 2006, we have collected the needed hardware for this task, but have not started speaker collection. We plan to have probe data completed by May 2006.

Emotion:

Speech with emotions such as anxiety, fear and anger is common when accessing automatic speech systems. For example, a person trying to access his bank account on his cell-phone might get frustrated due to repeated failures of the voice-based security system. Speech under emotion will also be a part of the proposed speech corpus (we have previously employed the Soldier of the Month paradigm [4] in our algorithm development for stress detection and assessment). Speech collection under the last three stress conditions (2.2, 2.3, 2.4) is currently in progress.

Evaluations

Some preliminary experiments were performed to investigate the effect of Lombard speech on system performance. Also, perceptual speaker ID experiments are presently being performed.

In-set Speaker ID tests:

An in-set speaker identification system classifies input speech as belonging/not belonging to a specific group of speakers defined in the system. The system, however, does not classify which speaker within the in-set group is actually speaking. The in-set speaker ID system uses Gaussian Mixture Models (GMM) to model each of the speakers in the given set. From the input test speech, the probability of the input being produced by the different speakers is computed and a decision is reached regarding the speech being in-set or out-of-set. Unconstrained Cohort Normalization with Likelihood Ratio Testing (UCN-LRT) is used to form decision rules. Further details of this system can be found in [12], [13].

A set of 30 speakers (19 female and 11 male) are selected, with 15 chosen as in-set speakers, 15 set aside as out-of-set speakers. The GMM's for each in-set speaker are trained using 30 seconds of neutral speech. The system is tested with speech from all the 30 speakers. Both neutral and Lombard speech are used for testing. Test utterances of 3 and 12 seconds are used in two different experiments. The equal error rate results are summarized in Tables 2 and 3.

Noise Type	Noise Level 1	Noise Level 2	Noise Level 3
HWY	29.83	35.5	39.17
LCR	33.67	34.33	35.5
PNK	27.83	32	35

Table 2: EER (%) of in-set speaker ID system for 3s Lombard speech. Noise levels 1,2,3 are[LCR,HWY: 70,80,90 dB-SPL; PNK:65,75,85 dB-SPL). EER for clean neutral speech: 9.833 %

Noise Type	Noise Level 1	Noise Level 2	Noise Level 3
HWY	27.99	32.83	38.33
LCR	30	29	30.5
PNK	21.17	27.67	31.7

Table 3: EER (%) of in-set speaker ID system for 12s Lombard speech. Noise levels 1,2,3 are[LCR,HWY: 70,80,90 dB-SPL; PNK:65,75,85 dB-SPL). EER for clean neutral speech: 3.67 %

From the above EER scores, we find that speaker ID performance deteriorates greatly under Lombard effect. Also, test utterance length has no effect on the EER of the system for Lombard speech, which is not the case for neutral speech, where increased duration test material improves performance.

CONCLUSIONS

In this paper, a speech corpus with speech influenced by human factors such as stress and emotion has been proposed. With speech consisting of spontaneous and read sentences, data from many subjects, recordings from different microphones, and multi-modal data (video recordings and biometrics), this corpus will be a rich resource for research on detection, analysis and compensation for emotion, stress and Lombard effect. In this preliminary stage, data collection is almost complete for Lombard effect, with initial collection started in other phases. Experiments using in-set speaker ID system show that Lombard effect degrades speech system performance. Thus, a useful research direction is to perform detailed analysis on this corpus to understand the science of speech production under Lombard effect and the specific differences in the characteristics of Lombard speech. Also, from the analyses, approaches to normalize Lombard speech for speech system performance improvement can be developed.

ACKOWLEDGEMENTS

We thank Prof. Linda Thibodeau from the School of Brain and Behavioral Sciences at Univ. of Texas at Dallas for making her laboratory and audiometry equipment available to us during the speech corpus collection.

REFERENCES

1. Hansen, J.H.L., Swail, C., South, A.J., Moore, R.K., Steeneken, H., Cupples, E.J., et.al, *The Impact of Speech Under `Stress' on Military Speech Technology,* NATO Research & Technology Organization RTO-TR-10, March 2000 (ISBN: 92-837-1027-4).
2. J.H.L Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition,* Ph.D. Thesis, School of Electrical Engineering, Georgia Institute of Technology, July 1988.
3. J.-C. Junqua, *The Lombard reflex and its role on human listeners and automatic speech recognizers,* Journ. of Acoustical Society of America, Jan. 1993.
4. E. Ruzanski, J.H.L Hansen, et.al, *Improved "TEO" feature-based automatic stress detection using physiological and acoustic speech sensors* , Interspeech 2005.
5. M. Akbacak, J.H.L Hansen, *Environmental sniffing: noise knowledge estimation for robust speech systems*, International Conference on Acoustic and Speech Signal Processing 2003.
6. J.H.L. Hansen, S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," *EUROSPEECH-97*, vol. 4, pp. 1743-1746, Rhodes, Greece, Sept. 1997
7. H. Abut, J.H.L. Hansen, K. Takeda, *DSP for In-Vehicle and Mobile Systems,* Springer-Verlag Publishing, Oct. 2004.
8. H. Abut, J.H.L. Hansen, K. Takeda, *Advances in DSP for In-Vehicle and Mobile Systems,* Springer Publishing, scheduled July 2006.
9. J.H.L. Hansen, X.X. Zhang, M. Akbacak, U.H.. Yapanel, B.Pellom, W. Ward, P. Angkititrakul, "CU-MOVE: Advanced In-Vehicle Speech Systems for Route Navigation," Chapter 2 in *DSP for In-Vehicle and Mobile Systems*, Springer-Verlag, 2004.
10. J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Comm.,* vol. 20(2), pp. 151-170, Nov. 1996.
11. J.H.L.Hansen,"Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," *IEEE Trans. on Speech & Audio Processing*, vol. 2, no. 4, pp. 598-614, Oct. 1994.
12. Angkititrakul et al., "Cluster-dependent Modeling and Confidence Measure Processing", ICSLP 2004.
13. Fortuna J., Sivakumaran P. et al. "Open-set speaker identification using adapted Gaussian mixture models", Interspeech 2005.

The SAFE Corpus: illustrating extreme emotions in dynamic situations.

Chloé Clavel

Thales Research & Technology France
RD 128, 91767 Palaiseau Cedex
Chloe.clavel@thalesgroup.com

Ioana Vasilescu, Laurence Devillers

LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France.
Ioana.vasilescu@limsi.fr, devil@limsi.fr

Gaël Richard

ENST-TSI
46 rue Barrault, 75634 Paris, Cedex 13, France
Gael.richard@enst.fr

Thibaut Ehrette, Célestin Sedogbo

Thales Research & Technology France
RD 128, 91767 Palaiseau Cedex

Thibaut.ehrette@thalesgroup.com

ABSTRACT

Existing real-life corpora illustrate everyday life contexts in which social emotions frequently occur. The type of emotional manifestations and the degree of intensity of such emotions are determined by politeness habits and cultural behaviours. This paper shows how the challenge of collecting extreme manifestations of emotion has been addressed with the acquisition of a corpus of fiction, the SAFE Corpus. The aimed application is civil safety and surveillance of public places in particular. A task-dependent annotation strategy is developed with both generic and specific descriptors. A description of the emotional content of the SAFE Corpus is provided. The corpus focuses on the illustration of extreme fear-type emotions in rich and varied contexts. Finally, a detection system of fear emotions based on acoustic cues has been developed to carry out an evaluation.

Keywords

Emotions in abnormal situation, civil safety, audiovisual corpus, annotation strategy, fear detection system.

INTRODUCTION

Emotions represent a complex research field with a large background in the communities of life and social sciences. In these contexts several theories have been proposed to define the concept "emotion" [5]. The emerging field of emotion analysis and detection in speech requires adapting these theories to the context of study and considering the variability inherent to the oral character.

In this paper, the aimed application is civil safety and surveillance of public places. Since current systems are mostly video-based, one of the main challenges is to use the audio content as complementary information to video to automatically detect an abnormal situation (situation during which the human life is in danger). The human oral communication in such situations is strongly based on the emotional channel. There is, as a consequence, a strong interest to automatically detect symptomatic emotions occurring in abnormal situations.

Studies dedicated to the analysis of emotion in speech commonly refer to a restricted number of emotions named "primary". The "Big Six" are, for example, frequently studied (fear, anger, sadness, joy, disgust and surprise). However emotions occurring, in everyday life contexts, are often a mix of several emotions or emotional manifestations which correspond more to the attitudes.

In this study, the targeted emotion is a primary emotion, namely *fear*. We are looking for fear-type emotions occurring in dynamic situations, during which the matter of survival is raised. In such situations the emotional manifestations correspond to primary manifestations of fear: they occur as a reaction to a threat. Their degree of intensity is particularly high and such emotions are rare in existing real-life corpora. Those corpora illustrate life contexts in which social emotions frequently occur. Abnormal situations are indeed rare and unpredictable and real-life recordings of abnormal situations are for the most confidential.

The lack of corpora illustrating strong emotions in real abnormal situations has encouraged us to build the SAFE Corpus (Situation Analysis in a Fictional and Emotional Corpus). The fiction provides an interesting range of potential real-life abnormal contexts and of type of speakers that would have been very difficult to collect in real-life.

The final goal is to develop a fear-type emotions detection system based on audio cues. The acoustic modeling of emotional speech obtained in such variable conditions is however more complex than for studies carried out on simulated emotions in laboratory conditions with a small number of speakers. The challenge is to control the variability of emotional manifestations by an appropriate annotation strategy. The annotation strategy has to be not only relevant for the application but also sufficiently *generic* to be exported to other corpora. In the next section, the SAFE Corpus is presented. The third section is dedicated to the annotation scheme according to which we provide a description of the corpus content in the fourth section. Finally, the last section gives some evaluation results of an automatic fear detection system using this corpus.

THE SAFE CORPUS

The SAFE Corpus consists of 400 audiovisual *sequences* from 8 seconds to 5 minutes extracted from a collection of 30 recent movies on DVD support. We focused for the sequence selection on the manifestation of emotional states in two contexts: normal vs. abnormal situations illustrated by individuals groups and/or crowds. Variability in terms of sequence duration depends on the way that situation is presented in the movie. Among the abnormal situations illustrated in the corpus we can mention natural damages such as fires, earthquakes, flood etc, physical or psychological threatening and aggression against human beings (kidnapping, hostages, etc.). A major contribution of such a corpus relies on the dynamic aspect of emotions: the corpus illustrates the emotion evolution according to the situation in interpersonal interactions. Besides, the fiction allows the collection of emotional data with their environmental noise. A total of 7 hours of recordings is thus collected in which speech represents 76% of the data. 71% of speech occurs in abnormal situations. The movies make use mostly of American English (70% of the data). In the remaining movies, actors are portraying other English (British, Irish, Canadian) or foreign accents (French, Scandinavian, German). The surveillance application implies to cope with a high number of unknown speakers. The SAFE Corpus provides about 400 different speakers in this purpose. The repartition of speech duration according to gender is as following: 47% male speakers, 32% female speakers, 1% child. The remaining 20% of spoken duration consists in overlaps between speakers, including oral manifestations of the crowd (2%).

ANNOTATING EMOTIONS IN DYNAMIC SITUATIONS

The annotation strategy takes into account the temporal aspect of the sequence. The emotion evolution is captured along the sequence by segmenting each sequence which provides a particular context into a basic annotation unit. This unit is called *segment* and corresponds to a speaker turn or a section of speaker turn portraying the same annotated emotion. The chosen descriptors are task-oriented. However the annotation strategy proposes also abstract descriptors which are more generic and can be exported to other applications. Annotation choices are helped by the video. A multimodal tool (ANVIL [7]) is used for the annotation. It is especially difficult to delimitate accurate emotional categories (in terms of perceived classes for the annotation strategy and of acoustic models for the detection system) when the data illustrate a large degree of diversity. To overcome these challenges, the annotation strategy has been developed with the consideration of various levels of accuracy. The segmentation and the annotation of the corpus were carried out by a first English native labeller. A second French/English bilingual labeller independently annotated the emotional content of the pre-segmented sequences. For these two labellers, the annotation of a given segment is influenced both by audio (acoustic and semantic content) and video information contained in the whole sequence. In order to evaluate the audio cues weight to detect a situation, which provokes fear, a supplementary “blind” annotation based on the audio support only have been done on a sub corpus [4]. This annotation is based on the listening to the segments in random order with no access to the contextual information conveyed by video and by the global content of the sequence.

Emotional Descriptors

The description of emotional substance is considered at the segment level and consists of two types of descriptors: dimensional and categorical. *Categorical descriptors* provide a task-dependant description of the emotional content with various level of genericity towards the corpus. We selected so far four major emotion classes: *global class fear*, *other negative emotions*, *neutral*, *positive emotions*. *Global class fear* corresponds to all fear-related emotional states. This last broad emotional category is completed by an oriented verbalization: the labeler has to precise the type of *fear* present in the segment by choosing a predefined (or not) sub-category (stress, terror, anxiety, etc.). *Dimensional descriptors* are based on the 3 abstract dimensions previously exploited in the literature [8]: activation, evaluation and control. The control dimension has been adapted here according to the application and renamed reactivity. The reactivity value indicates whether the speaker seems to be subjected to the situation (passive) or to react to it (active). Abstract dimensions are evaluated on discrete scales. The perceptual salience of those dimensional descriptors was evaluated in a former study [1]. Abstract dimensions allow specifying the broad emotional categories by combining the different levels of the scaled abstract dimensions.

Context Descriptors

The context of emotion emergence is described by a threat and a speaker track. The speaker track mentions the gender of the current speaker and its position in the interaction (*victim* or *aggressor*). The threat track provides the description of the threat intensity (4 levels scale) and of its incidence (*potential, latent, immediate, passed*).

Audio and Verbal Context Descriptors

Extracted sequences provide recordings made in variable conditions. Consequently, quality has a high variability inter- and intra-sequences. Labels defining audio quality of each segment are stored. We evaluate both the perceptual quality and we annotate the acoustic events which determine this quality. More precisely, the quality is estimated on a 4 steps scale: from *Q0* (inaudible speech or with too much sound effects) to *Q3* (perfectly intelligible speech and realistic sound recordings). The presence of noise and/or music is besides annotated according to 4 recording conditions: *Clean* (segments without noise and music), *Noise only*, *Music only* and *Noise and music*. The verbal and non verbal (cries, breathing, etc.) contents of the segments are also transcribed.

EMOTIONAL CONTENT

A total of 4073 speech segments with a duration ranging from 40 ms to 80s are obtained from the corpus sequences. The emotional content is presented in this section by considering the percentage of attributions for each label by the two labelers, so that the two annotations are taken into account. The inter-labeller agreement for the emotional content has been evaluated in another study [4]. The percentage of attributions of the four emotional categories is thus the following: 29% for *fear*, 30% for *other negative emotions*, 33% for *neutral*, and 8% for *positive emotions*. In this section we focus on the main features characterizing the emotional content: the presence of extreme fear as illustrated by abstract dimension intensity and the relationship between the emotion label and context (threat and environmental noise). A complete description of the corpus has been provided in a previous study [2].

Presence of Extreme Fear

The combination between categories and dimensions allows a best visibility of the types of manifestations contained by each category.

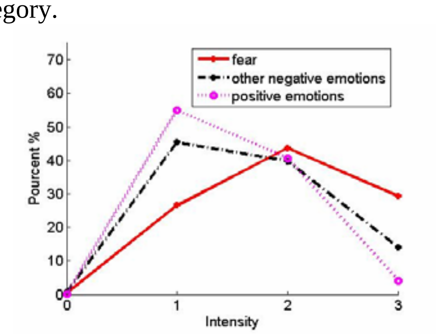


Figure 1. Distribution of each emotion category according to the intensity scale

Figure 1 shows the distribution of the attributions of each emotion category according to the first dimension, *intensity*. Fear-type emotions are perceived as more intense than other emotions. 73% of *fear* segments are labelled as level 2 or 3 on the intensity scale while the major part of other emotions are labelled level 1. Besides, the presence of cries (139) seems to be associated with the presence of extreme fear.

Emotions in Dynamic Situations

The correlation of categorical descriptions of emotions with the threat provides a rich material to analyze the various emotional reactions to a situation. Table 1 shows the distribution of each emotional category *(fear, other negative emotions, neutral, positive emotions*) as a function of the threat incidence. 96% of *fear* segments occur in abnormal situations and 53% when the threat is immediate. We can notice the presence of other emotions than *fear* in abnormal situations. 61% of the segments labeled *other negative emotions* emerge during latent threats (30%) or immediate threats (31%).

	Abnormal situations/Threat				Normal/No Threat
	Potential	Latent	Immediate	Passed	None
Fear (29%)	4	32	**53**	7	**4**
Neg. (30%)	9	30	31	6	24
Neu. (33%)	7	25	14	7	47
Pos. (8%)	2	10	4	5	79

Table 1. Percentage of types of threat per emotional categories

Figure 2 shows the distribution of each intensity level inside the *fear* class according to the threat intensity. According to the table 1, 4% of segments labeled *fear* emerges in normal situations (No threat). It emerges from figure 2 that these segments (Threat Intensity = 0) corresponds to types of fear with a low level of emotional intensity such as anxiety or worry. 13% of *fear* segments labeled level 1 on the intensity scale occur in normal situation vs. 0.5% of *fear* segments labeled level 3. The segments labeled level 3 on the intensity scale correspond for the major part (57%) to threats with a higher level on the intensity scale. Besides *Fear* segments occurring when the threat is labeled with a low level of intensity are essentially labeled level 1 on the intensity scale

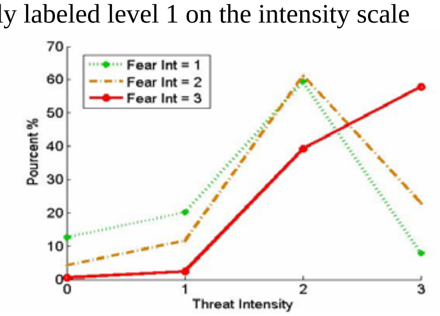


Figure 2. Emotional intensity of fear-type emotions according to threat intensity

Emotions in their Environmental Noise

Figure 3 shows that negative emotions correspond to noisier contexts than positive and neutral and contain more sound effects. In most movies, recording conditions tend to mirror reality: speaker movements implying natural variation in voice sound level are thus respected. However, the principal speaker will be more often audible in the fiction context. We can hypothesize that this is not systematically the case in real recording conditions. The categories obtained via this annotation could be employed to the test of the robustness of detection methods to the environmental noise.

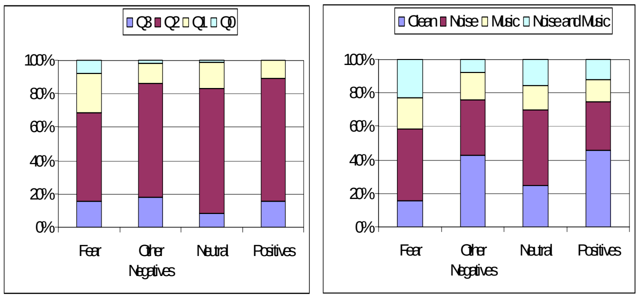


Figure 3. Emotional Categories vs. level of quality (left) and environment type (right)

EVALUATION RESULTS

We present here the first evaluation results of the fear-detection system developed on this corpus. The fear detection system is based on acoustic cues and focus as a first step on a fear vs. neutral classification of the emotional *segments*.

Experimental Database

The detection system is performed on a sub corpus containing only good quality (*Q2* or *Q3*) segments labeled *fear* and *neutral*. Overlaps have been avoided. Only segments where the two human coders agree are considered, i.e. a total of 986 segments (606 for *neutral* and 380 for *fear*). The gender repartition of the neutral segments is: 68% of male speakers, 30% of female speakers, 2% of child. The gender repartition of the fear segments is: 34% of male speakers, 62% of female speakers, 4% of child.

Fear versus neutral classifier

The classification system presented in details in [3] merges two classifiers, *the voiced classifier* and the *unvoiced classifier* which consider respectively the voiced portions and the unvoiced portions of the segment. The first step of the overall system aims at extracting prosodic and voice quality features and the second step at reducing the feature space using the fisher selection algorithm and the Principal Component Analysis. The third step consists in the training of the models of the two classes for each voicing condition (using Gaussian Mixture Models or GMM). The final step consists in the classification of each segment according to the two main classes (the *fear* class and the *neutral* class) merging the results of the two classifiers (*voiced* or *unvoiced*).

Test Protocol

The test protocol is the protocol *Leave One Movie Out*: the data are divided into 30 subsets; each subset contains all the segments of a movie. The protocol consists in training the model on 29 subsets, leaving the last subset for testing and in iterating the procedure for the 30 possible test subsets. This protocol ensures that the speaker used for the test is not found in the training database. Detection performances are evaluated by the equal error rate (EER). The EER corresponds to the error rate value occurring when the decision threshold of the GMM classifier is set so that the recall will be approximately equal to the precision. The corresponding chance performances are thus 50%.

Results

Best results (EER = 30.5%) are obtained when the unvoiced classifier is considered with a weight decreasing quickly when the voiced rate increases [3].

CONCLUSION & PERSPECTIVES

This paper is dedicated to the presentation of the SAFE Corpus. The challenge taken up by this corpus relies on the illustration of extreme fear-type emotions in threat dynamic contexts. A first *fear* vs. *neutral* detection system has been performed on this corpus. The EER obtained by this detection system is reaching 30.5%. Further work will be dedicated to the evaluation of a *fear* vs. *other emotions* detection system and to the building of specific acoustic models which takes into account the variability of *fear* manifestations.

REFERENCES

1. Clavel C., Vasilescu I, Devillers L., Ehrette T., Fiction database for emotion detection in abnormal situation, ICSLP 2004, Jeju.
2. Clavel C., Vasilescu I., Richard G. and Devillers L. Du corpus émotionnel au système de detection : le point de vue applicatif de la surveillance dans les lieux publics. Accepted for publication in the French Revue in Artificial Intelligence (RIA) (2006).
3. Clavel C., Vasilescu I., Richard G. and Devillers L. Voiced and Unvoiced content of fear-type emotions in the SAFE Corpus. Speech Prosody 2006, Dresden – to be published.
4. Clavel C., Vasilescu I., Devillers L., Ehrette T. and Richard G., SAFE Corpus: fear-type emotions detection for surveillance application, In Proc. International conference on Language Resources and Evaluation, Genoa, Italy, May 2006 – *to be published.*
5. Cowie R., Cornelius R., Describing the emotional states that are expressed in speech, Speech Communication – Speech and Emotion, volume 40, pages 5-32 (2003).
6. Douglas-Cowie, E., Campbell, N, Cowie R., Roach R., (2003), Emotional speech: Towards a new generation of databases, Speech Communication, vol 40, pages 33-60.
7. Kipp, M., (2001). Anvil a generic annotation tool for multimodal dialogue. Eurospeech.
8. Osgood, C., May, W.H., Miron M. S. (1975), Cross-cultural universals of affective meaning, University of Illinois Press, Urbana.

A Speech Corpus with Emotions

Janne Bondi Johannessen, Kristin Hagen, Joel Priestley and Lars Nygaard

The Text Laboratory, University of Oslo
P.O. Box 1102 Blindern, 0317 Oslo, Norway
{jannebj, kristiha, joeljp, larsnyg}@iln.uio.no

ABSTRACT

The language data in most language corpora are gathered under extremely controlled conditions, missing out a large array of facts about human language, especially language in emotional settings. In order to make up for the usual shortcomings, we have developed a speech corpus that consists of two parts; one with footage from the TV show Big Brother, and one with interviews with informants plus dialogues between pairs of informants. The first part has a lot of very emotional situations involving the Big Brother participants in many real-life situations, while the second part has informants with different relations, such as parent-child, friends etc. The corpus is transcribed, grammatically annotated and searchable in a variety of ways via a very friendly user-interface. Importantly, the transcriptions are linked to sound and video. This paper gives an overview of the advantages of the corpus, and provides examples of emotional speech situations, with some linguistic characteristics.

Keywords

Speech corpus, emotions, Norwegian, Big Brother, conversations, interviews, informants, transcriptions linked to video files.

INTRODUCTION: THE NOTA CORPUS

In this paper, we shall show how a speech corpus consisting of more than one kind of speech situations can fulfill many of the needs that linguistics and language technology may have. We will focus on the NoTa Norwegian Speech Corpus [4] of 2 million words, in which we have sought to avoid some of the shortcomings of other written and spoken language corpora. We have had these goals when building the corpus: 1) to include multimedia representation (transcriptions of speech, linked to audio and video representations); 2) to include a variety of speech situations, especially emotional ones; 3) to make a representative selection of informants; 4) annotate the corpus in various ways (relating to emotions, dialogue, grammar, extra-linguistic features); 5) to make new annotation possible; 6) to use cost-effective methods; 7) to obide by legal rules.

The NoTa corpus consists of two parts at the moment: NoTa-Oslo and Big Brother, but it will consist of more parts and is growing all the time. At the moment, the additional parts that are planned and are at various stages of development are TAUS (a thirty year old speech corpus with tapes and transcriptions), UPUS (a sociolinguistic corpus of young people's speech) and ScanDiaSyn (a dialect corpus with tapes and transcriptions from the whole Scandinavian dialect area).

Here, we will focus on the two subcorpora Big Brother and NoTa-Oslo, both of which are searchable w.r.t. strings and the latter also by grammatical tags, and are available on the Internet for research via a fully functional user-interface that includes sound and video. Section 2 discusses the background for the corpus, while section 3 describes the contents of the two subcorpora. In section 4, examples of emotional situations from the corpus are shown and characterised linguistically. Section 5 contains some technical information.

BACKGROUND

The use of text corpora for the studying of language has gained wide support over the last couple of decades, whether the aim is a proper understanding of intricate linguistic structures or quantitative or statistical knowledge for NLP purposes. However, they usually contain only written or spoken language gathered under extremely controlled conditions, and lack a wide range of central linguistic features: all kinds of features (structural and lexical) relating to dialogue, such as interruptions, repetitions, feedback etc., and especially features relating to emotional situations, such as grammatical constructions like imperatives and exclamatives, and phonological and lexical features.

Some spoken language corpora have more variation. The British National Corpus [2] and the Gothenburg Spoken Language Corpus [3] both contain a variety of speech situations, including some emotional ones. However, both have a serious disadvantage. They do not have multimedia representations of the data, only written transcriptions of the speech.

The linguistic expressions of emotions have not been widely studied by linguists. While language technology, and in particular the speech synthesis domain, has been interested in the topic for some time [6], [8], and psychologists have been interested in the emotional situations, linguists have not yet discovered the wealth of data that emotional situations offer. A properly annotated corpus with a user-friendly interface may change that, making the corpus truly multi-purpose.

INFORMANTS AND SITUATIONS IN THE CORPUS

Problems with Legislation

According to Norwegian legislation, it is necessary to obtain written consent from every informant. This rules out any natural setting in which subjects would unknowingly be taking part. We found two ways around this, both of them naturalistic: a) Using a reality show where the subjects had already given up all personal rights. b) Using pairs of hand-picked informants.

The Big Brother Subcorpus

We chose the first season of the television show Big Brother. In contrast to the participants in the later series, the first ones were not overtly exhibitionistic or playing to the TV audience. The show is full of situations with most of the basic emotions suggested by Robert Plutchik [7]: Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise. It is very useful for linguistic and language technological purposes, but has also been used by psychologists and sociologists for studies on human behaviour more generally. A problem with it is that the range of informants is very narrow: only young adults in their twenties and thirties without higher education. Another problem is that the director of the show compiled footage in a way that might give a false picture of the situation.

The NoTa-Oslo Subcorpus

The NoTa-Oslo subcorpus has pairs of informants. We have used nearly 170 informants, carefully chosen w.r.t. sociolinguistic variables (age, gender, education, place of dwelling, of birth, and of childhood, occupation, but not name). The people have different relationships to each other: parent-child, spouses, friends, colleagues, acquaintances and unknown. The informants take part in two different kinds of situations: a semi-formal interview with the RA and a comfortable, laid-back conversation (dialogue) with one other informant. These two situations have been maximally differentiated, so that during the conversation, the informants are given sweets, drinks, and the table is covered by an attractive table cloth, mugs from an assorted collection etc. The two situations are perceived as very different by the informants, witnessed by very different kinds of language (hesitations and pauses vs. fluent speech and interruptions) and actions such as fits of laughter, falling off the chair, flirtation etc. in the conversation part. Some advantages of this part of the corpus are that we know a lot about the informants; that we have each of them in two comparable situations; and that the informants represent a wide range of people. But it is also clear that the range of emotions is not as wide as with the Big Brother part. Furthermore, the Norwegian legislation require a high level of anonymity and security (to the extent that this is possible when they appear on sound and video). This has two consequences. First, the topics that are talked about must be "safe": the informants must be instructed not to talk about e.g. politics, religion, illness, criminality, and other people. Second, the informants must not be linked to the data by name or other identification, so the lists of their names and addresses have had to be destroyed.

EMOTIONS IN THE CORPUS

At the moment, the corpus is annotated grammatically, and in addition, a range of extra-linguistic noises, such as laughter, coughing, sucking noises, smacking noises etc. are marked up. The corpus is not yet annotated with emotions. However, many emotions are usually accompanied by certain words or expressions, which could be said to identify them, and since the corpus is transcribed, it is possible to search for these. Since the transcriptions are linked to sound and video, it is easy to check the emotions in question. As expected, there are more dramatic emotions in the Big Brother subcorpus than in the NoTa-Oslo one. Together, however, they represent a wide range of emotions. Out of the eight emotions Plutchik's basic emotions, mentioned above, there is probably only one that can be hard to find in the corpus, and that is fear. While the informants may discuss fear, they are obviously not necessarily feeling it at the same time. Below we will show some examples of emotional situations from the corpus, and characterise them linguistically.

Anger

There is a lot of anger in the Big Brother corpus. The people in the Big Brother house are going through many ordeals, including such that arise from them being young and full of feelings. The utterance below is uttered by a young woman as she is going into the shower, and she is annoyed with one of the young men, who is teasing her by standing naked just in front of her:

> **Dø, kutt ut a for faen. Nå er jeg trøtt og nå er jeg sur og nå sa jeg det.**
> *(you cut out then for devil now am I tired and now am I sour and now said I it)*
> *('Hey f.. you, cut it. Now I'm tired, and now I'm annoyed, and now I've said it.')*

Example 1: Anger (uttered by Rebekka, 27).

Finding this and other examples of angry emotions in the corpus can be done simply by searching for the swear word *faen*. Studying the utterance more carefully, we find several linguistic features that exhibit angry emotions apart from the actual lexical contents that describe the emotions. First, phonologically, the utterance has falling intonation, contrary to how more neutral emotions are expressed in Norwegian. Second, there is heavy stress on several of the words (each time the word 'now' is uttered, it is maximally stressed). Third, there is repetition and quasi-repetition of whole phrases (Twice: *'Now I am X'. Once: Now I said X*). Fourth, there is the swear word phrase *for faen* that is mostly used in angry contexts. Fifth, the imperative sentence is initiated by addressing the listener with an impolite version of the second person pronoun, with no other function than emphasising the anger. In addition the many features about angry speech we find in this little extract, it is interesting to note that the language uttered spontaneously by this angry person has a high degree of rhetoric quality. The repeated phrases have a nice melodic metric that makes quite an impact on the listener, and the three repeated phrases that twice differ only with respect to the predicative adjective and once with respect to the whole predicate, have a much stronger impact power than if she had avoided repetition, and simply coordinated the three different predicates.

Surprise

The NoTa-Oslo corpus is full of surprise contexts. This is to be expected when two people are sitting down to talk together for 30-40 minutes. One informant often ends up telling the other things that the other had either mistakenly believed not to be the case, or would never have thought possible of that person or in that situation. Below, a young girl twice expresses surprise that the person she talks to did not sleep in a hotel.

090 jeg sov bak i bussen
(I slept behind in the bus)
089 hadde du ikke hotell?
(Did you not have a hotel room?)
090 nei # jeg sov sånn derre litt her og litt der og
(No, I slept a bit here and a bit there)
089 hæ jeg trodde du hadde hotell jeg
(what I thought you had hotel I)
('What! I thought you had a hotel room!')

Example 2: Surprise (uttered by 089: a young girl, 18, talking to 090: young man, 18).

Finding situations of surprise is not difficult. The informal word *hæ*, possibly derived from *hva* 'what', typically characterizes surprise, so it can be used as a search key word. The little situation in the example shows two utterances of surprise. The first utterance is a question expressing not just the speaker's wish to find out what happened, but also reveals disbelief at the claim. The full meaning of the question is expressed mainly by the intonation, which is not simply rising, as in other questions, but is actually lowered again at the last syllable. The second utterance of surprise is characterised by a number of features. The exclamative *hæ* is the first and most obvious one. Second, the intonation here, too, plays a role, being more vivid than if the utterance was simply one of a more neutral statement. Third, the utterance contains a copy of the subject in a right dislocated position, again adding an extra bit of emphasis to the utterance.

Acceptance

The two subcorpora differ substantially w.r.t. situations of acceptance. It is fair to say that NoTa-Oslo contains nearly only situations of acceptance, given the nature of conversations where the informants are knowingly being taped. In example 3, two young girls are having a very friendly conversation, where one is eagerly telling the other about how she got to see all the participants in a costumes ball because she was working at the entrance, while the other is eagerly listening, expressing interest and consent.

007 og da så jeg alle som kom inn # m da
da fikk jeg med meg ganske mange kostymer
(And then I saw everybody who wanted to come in - I got to see quite a few costumes.)
008 * mm
(I agree.)
007 jeg syns det var morsomt med de guttene som var gresskar jeg # for de var sånn seks stykker eller sånn
(I really liked the blokes that were pumpkins)
008* mm
(I agree.)

Example 3: Acceptance (uttered by 007: a young girl, 17, talking to 008: young woman at 18).

Since the corpus is full of feelings of acceptance, it is not difficult to find instances of it, but it is also possible to search for a typical linguistic expression, such as the disyllabic exclamation *mm* 'I agree' (not to be confused with other nasal exclamations).

Affection

Affection is not one of Plutchik's basic emotions, but it is often included in lists of emotions. Affection exists in the Big Brother corpus, but more surprisingly, it can also be found linguistically expressed in the NoTa-Oslo part:

132 [laughing-] man må ta hensyn vet du [-laughing]
(You have to be considerate, you know)
131 [laughter]
132 kan oppstå situasjoner forstår du #
(Situations can arise, you see)
men nå skal ikke du tigge # nei ## [laughter]
(But you mustn't beg, no)
[laughing-] åssen går det med [-laughing] katta di da?
(How's your cat?)

Example 4: Affection (uttered by 132: a woman, 52, to dog while talking to 131: a woman, 31).

Some of the dialogues in NoTa-Oslo involve people who have brought their dogs or babies with them, and it is in these contexts that we find loving expressions. In example 4, two colleagues are talking casually, when one of them suddenly addresses her dog. The grammatical structure of her utterance is not different from the rest of the dialogue, but her voice becomes lower, both in pitch and volume, and takes on the affectionate accent of "motherese". She then forgets the old conversation topic and goes on to ask her colleague about her cat.

TECHNICAL INFORMATION

Both parts of the corpus are represented with video and audio recordings linked to transcriptions of all speech. We have used Quicktime Pro to convert from .wav-format to AAC in .mov-files, to be played by each user in Quicktime, via a central streamer. All speech is transcribed using the free program Transcriber. The corpus is tagged by a HMM tagger trained on a manually corrected version of the Oslo-Bergen tagger.

The corpus is searchable via the Internet page using a very user-friendly interface built on top of the IMS Corpus Work Bench Query system. All the transcriptions of the speech occurring in the corpus are searchable, as are the specially annotated events such as laughter and coughing, plus a variety of interjections and exclamations, extra-linguistic noises etc. It is also possible to do searches via grammatical tags. The search results are presented as concordances that are clickable for sound and video as well. The corpus can also be downloaded to see the full transcriptions and view and listen to the full recordings.

CONCLUSION

Focussing on two subcorpora of speech recorded in very different situations, and presented in a user-friendly search-interface with transcriptions linked to sound and video, we have shown that it is possible to get good data for linguistic research on emotions.

ACKNOWLEDGEMENTS

We want to thank research assistant Anne Marit Bødal and transcribers Hilde Cathrine Haug, Ingunn Indrebø Ims, Signe Laake, Inger Margrethe Hvenekilde Seim, and Åshild Søfteland for providing us with valuable information on specific parts of the corpus. We also would like to thank Kristin Eide and Ingunn Amdal for useful comments.

REFERENCES

1 IMS Corpus Work Bench:

http://www.ims.uni- stuttgart.de/projekte/Corpus Workbench/OldDocus/FAQ.html

2.The British National Corpus:

http://www.natcorp.ox.ac.uk/

3. The Gothenburg Spoken Language Corpus:

http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3

4. The NoTa Corpus:

http://www.hf.uio.no/tekstlab/nota/english/index.html

5.Transcriber:

http://trans.sourceforge.net/en/presentation.php

6. Eide, E., A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli. 2004. A Corpus-Based Approach to Expressive Speech Synthesis. In *SSW5-2004*, 79-84.

7. Plutchik, Robert. 1980. *Emotion: a psychoevolutionary synthesis*. New York: Harper & Row.

8. Schröder, Marc. 2001. Emotional Speech Synthesis: A Review. *Eurospeech-2001*, 561-564.

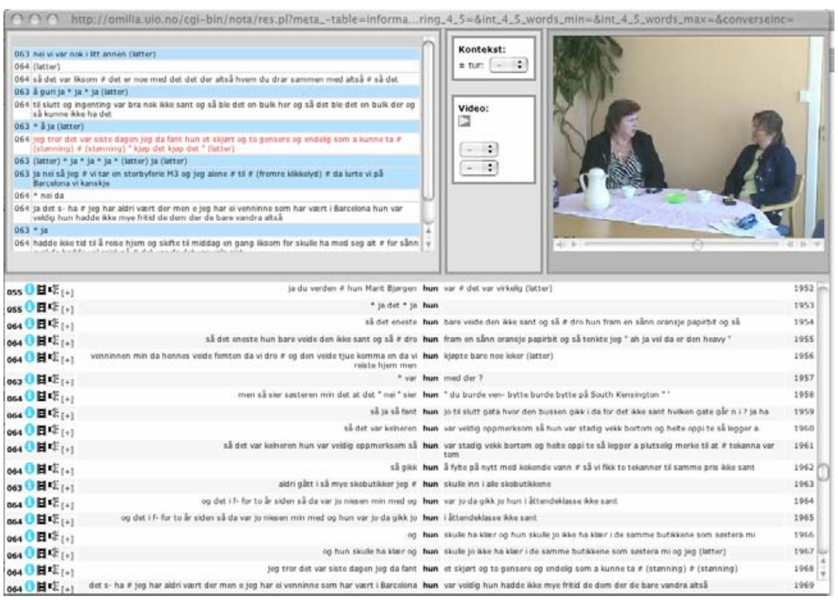


Figure 1: Search results with video recordings clickable from concordance lines.

Annotating State of Mind in Meeting Data

Dirk Heylen *Dennis Reidsma* *Roeland Ordelman*

University of Twente
Human Media Interaction
{heylen,dennisr,ordelman}@ewi.utwente.nl

ABSTRACT

We discuss the annotation procedure for mental state and emotion that is under development for the AMI (Augmented Multiparty Interaction) corpus. The categories that were found to be most appropriate relate not only to emotions but also to (meta-)cognitive states and interpersonal variables. The history of the development of the annotation scheme is briefly described. The discussion centers around the presentation of the procedure.

Keywords

Annotation procedure, Emotion and Mental State, Meetings

INTRODUCTION

The AMI corpus (see www.amiproject.org) is a collection of multi-modal meeting recordings. The majority of meetings were elicited using a scenario whereby groups of four participants played different roles in a corporate design team. More than one hundred hours of meeting data have been collected. The goals of the AMI project are manifold. The prime goal is to develop several kinds of "meeting technologies". The technology around which much of the research is centered in the first phase of the project is called the "Meeting Browser". The Meeting Browser is a collection of programs that allows people to have access to the recordings that were made. It involves special techniques in multimedia indexing, multimedia retrieval and multimedia extraction. The major effort at the first stage of the project was to acquire the data. The second stage involved defining and testing annotation schemes for various relevant dimensions for the various applications that will make use of the data collection. Currently, the main work is to annotate the recorded meetings manually with all kinds of information that can be used as meta-data for the recordings or, more importantly, for use by machine learning techniques that will automatically extract features from the data or that will generate annotations automatically. The recordings are of interest not just to the signal processing researchers, or the researchers dealing with multimedia information retrieval/extraction but also for people interested in face-to-face conversation for its own sake: conversational analysts, linguists, social psychologists etcetera.

Procedures that can determine what people feel and think when they are engaged in conversation can help us to retrieve information from recordings of people in a meeting. From the point of view of the relevance for meeting browsing and other techniques for building up memories of what happened in a meeting, it is obvious that what is relevant about what goes on in people's minds is not only what they "felt" about what was being said in the emotional meaning of the word, but also whether they were surprised by the things that were said, certain, sceptical or how clear or confusing certain issues were presented.

In the following sections, we discuss in detail our approach to the annotation for mental state for the AMI corpus. The main discussion centers around the presentation of the procedure. This involves segmenting the video and audio streams first and then assigning a number of features and categories to each of these segments. The annotation labeling consists of assigning categorical labels as well as indications of dimensional parameters (intensity and valence). We will discuss the selection of the categories. In passing, the history of the development of the annotation scheme is briefly described.

ANNOTATION PROCEDURE

The job of the annotator is to watch videos of the AMI meetings and annotate them with information about what we will call the "mental state" of the participants. This means that the annotator continuously tries to answer the question: "What state of mind is this person in?" Is the person happy, surprised, interested, bored?

There are two important points to make at the start. First, the notion "mental state of a person" can be loosely interpreted as "the feeling of a person". We want to stress that we consider feeling in a broad sense to include not only typical emotional categories such as "irritated" or "amused" but also so-called (meta-)cognitive states and processes such as "trying to remember", "paying attention" "interest", "distraction" or "puzzlement" and what could be termed propositional attitudes such as "scepticism" or "uncertainty". Another kind of category that was used included terms such as "dominance", "defensiveness" and "support" that provide a characterisation of the interpersonal variables.

A second point to note is that we tell annotators that we do not expect them to be able to completely mind-read what is going on with a participant. We discourage them from psychoanalysing a person in depth. Instead we ask them to give a broad description only of what they can reasonably

assume to be the mental state of the person, based on what they observe.
The instructions for the annotator and the whole procedure were developed after a number of sessions in which the developers of the scheme (i.e. the authors of this paper) watched and analysed several meetings together and individually; trying to achieve a consensus agreement on segmentation and labelling.
For each annotation assignment, annotators watch the video recordings of a meeting. They have a choice as to which viewpoints they want to watch. In particular, there are close-up recordings available for every single participant along with several overview videos that allow one to see the behaviour of the participant in context of the other participants. There are videos that show the pairs of meeting participants sitting at the same side of the tables and a video track, shot from the side that shows all participants. The close-up recording is generally used for the annotation of mental state, often accompanied by an overview video that provides more information about the context. The annotators can choose which videos they prefer to have open for inspection.
The annotation task consists of two parts: first, defining "cuts" (segmentation points) in the video of a person at places where a *distinctive change* in the mental state of this person occurs, and second, to fill in a form that describes each segment that is thus created. In brief, the instructions to the annotator read as follows:

- Start watching the video and try to imagine what the mental state is of the person you are observing.
- As soon as you notice a distinctive change in the mental state you press a key that will mark a segment boundary. The video stops playing. Note that this boundary will in fact be an "end-point". The start of this particular segment is the end-point of the previous segment (or in case it is the first segment, the start will obviously be the start of the video).
- You fill in a form that describes the mental state of the participant in the segment. This will include specifying the intensity, and the quality or evaluation of the mental state (whether it is a positive or a negative one). This will be explained in more detail below. You also have to choose one or more relevant category words from a predefined set that fits the mental state of the participant in the segment best.
- You press the "continue" key. The video resumes playing at the beginning of the next segment and the annotation process reiterates.

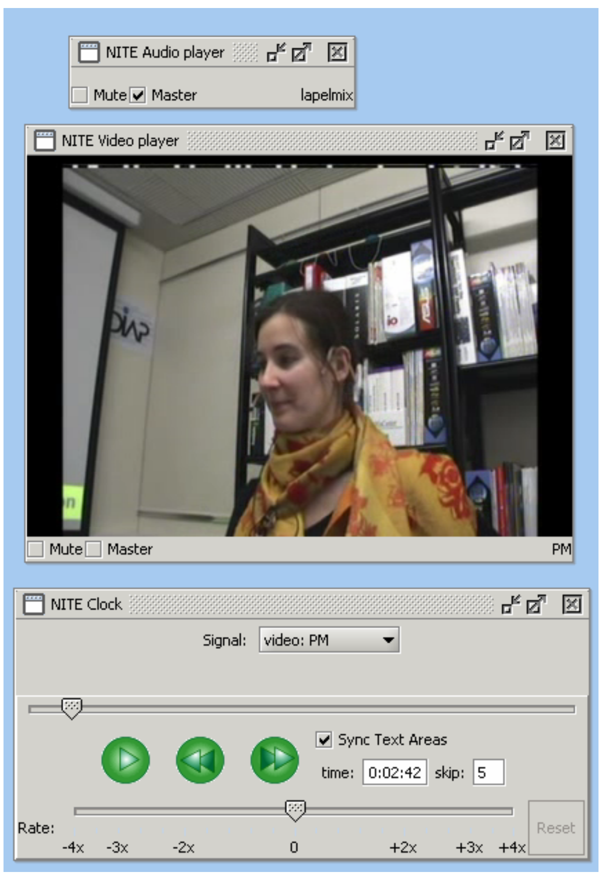


Figure 1 The video and audio controls

The video and audio controls are shown in Figure 1. The controls for the annotation form are shown in Figure 2.

CHANGES IN MENTAL STATE

The basis of the annotation process is marking up changes in the mental state of the participant in a video. These define the segments for which labels are defined. There are two types of change that we want to be annotated and that consequently should lead to the creation of a segment boundary: a change in mental state type and a clear change in the intensity of the mental state. Being amused, annoyed, angry, happy or relaxed are examples of mental state types. We explicitly defined one special "mental state": the neutral state. Evidently, a neutral mental state is a construct that does not really exist, but we define it here as an observed mental state that does not have a distinctive type with a particular intensity level. Such a segment will be denoted as a neutral segment with intensity 0.
A change in mental state type will be observed for example when somebody has a neutral look for a while, and one can observe that this person starts to look amused. Or if somebody has been looking very annoyed for a while, and then relaxes again to a neutral state. Typically the changes in mental states have longer or shorter fade-ins and fade-outs. A look of surprise may arise suddenly and disappear quickly, whereas amusement might start with a slight pulling up of the corners of the mouth (almost unnoticeable) that gradually builds into a complete smile and then slowly dissipates again. A clear change in the intensity of a mental state can be observed for example when somebody has been looking vaguely annoyed for a while, and suddenly the person starts to look extremely annoyed and frustrated. In this example, the annotator is asked to place a segment boundary when the intensity level starts to increase. The segment to the left of that boundary receives a type description "annoyed" with a relatively low intensity, and the segment to the right of the boundary a type description "annoyed" with the highest intensity value. Another typical example of an intensity change occurs when someone first looks amused and then starts laughing. The segment boundary should be placed just before the laughing starts. There are many cases where it is difficult to

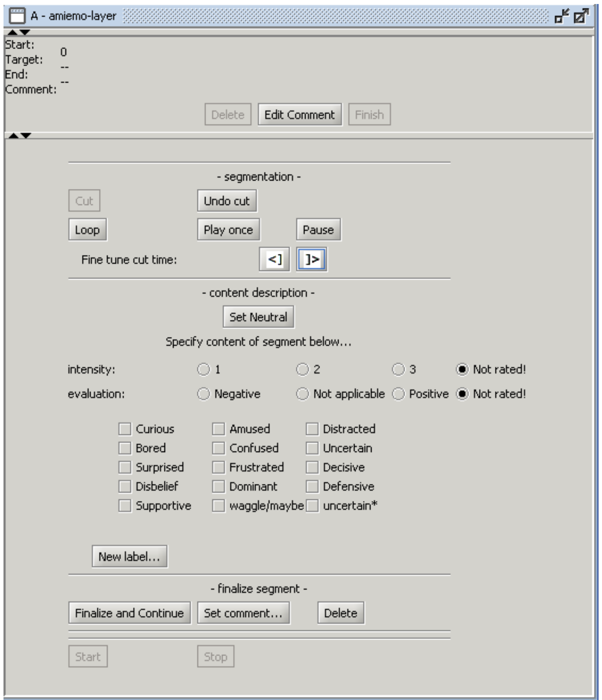


Figure 2 The emotion annotation interface

decide whether a type or intensity change should be marked as a segment boundary or not. In the example of the person that becomes amused (a smile develops, a smile remains for a while and gradually disappears) the choice will usually be to interpret this event as a single segment or "episode" which covers all three phases, instead of marking this as three separate segments differing in intensity. However, when the smile turns into a roaring laughter, the smile and the laughter should definitely be marked as two separate segments with different intensity values. In order to get familiar with these kinds of decisions, the instruction demos that are used for training the annotators contain segments with displays of the kinds of fluctuations in the intensity of certain mental states that we consider to be minor differences that do not deserve further segmentation. The instruction demos were created by the developers of the scheme.

Annotation dimensions

After having created a segment boundary, the mental state in the segment of focus has to be described. Forms[1] were designed for the annotator to make it easy to rate the intensity and the valence of the mental state and to choose a category label from a list of predefined labels that the annotator thinks best fits the observed mental state of the meeting participant in the segment of focus.

[1] The modules and tools were developed using the Nite XML Toolkit, an open source toolkit downloadable from http://www.sourceforge.net/projects/nite.

Intensity: Intensity is rated on a three-point scale: (1) low intensity, (2) medium intensity and (3) high intensity. A neutral mental state will automatically receive the intensity level 0.

Evaluation: People constantly evaluate the events, their own and others actions in many ways. This can be judgements about they are good or bad in a moral, ethical sense; whether they are good or bad for the goals they are pursuing; whether what is being said is true or false, believable or unbelievable, etcetera. Annotators are asked to mark the evaluation of a segment 'negative' in case the negative aspects of evaluation seem predominant for the participant, and positive otherwise. Sometimes positive and negative are not applicable, in which case one can choose the "not applicable" option. 'Positive' might for example be appropriate when someone is laughing; 'negative' for example when someone is very angry.

Category labels: Finally, a label that best describes the mental state of the participant in the segment has to be chosen from a predefined list. More than one label can be chosen, if the annotator feels that the segment fits into multiple categories. However, the annotators are told to do this only if there is not possible to assign one dominant state or if multiple mental states are clearly marked and deserve to be labelled as such.

The labels indicate quite general, diffuse categories. The names on the form have been chosen to indicate a category of labels rather than a specific state. For instance, the label "curious" should be used if one notices that the person observed shows a special interest in something. Other words that one might have used are "attentive" or "focused" or "interested". If the annotator thinks another label or description is exactly to the point given a particular observed mental state, there is room to add this word or description. Also, if the annotator feels that the segment does not fit any of the categories indexed by the labels, he or she can propose a new label and add this. We encourage annotators to use the standard label set, that was defined especially for the meeting domain, as much as possible.

Labels

Each of the labels that are currently in use is presented below. These labels were introduced by looking at several fragments of the data by the designers of the schema and by other annotators during a number of trials of the software and the schema. As we just said, some of the labels that were nearly synonymous were collapsed into one category.

Neutral: nothing remarkable is happening.

Curious: (interested, attentive, focused) The participant shows special interest in a topic or issue. In many situations "paying attention" is the neutral state (people are listening to others, for instance). These cases are labeled as neutral.

Amused (cheerful, joking): The participant is clearly amused.

Distracted (inattentive): The participant is not paying attention to the central issue in the meeting. The participant can be distracted by specific other things, or the participants mind may be wandering.

Bored: The participant is clearly bored with the proceedings of the meeting.
Confused (puzzled): There is something that the participant does not understand or that the participant cannot work out.
Uncertain (hesitant): The participant is not certain about something. ("I don't know")
Surprised: Something occurs or is said that the participant had not expected.
Frustrated (annoyed): The participant appears frustrated or annoyed about something.
Decisive (certain, confident): The participant is decisive, or very confident and certain about something and shows this by being more assertive and resolute than normal. This may be about an issue in the meeting (e.g. giving an opinion of which the participant is very certain), or about the meeting process itself.
Disbelief (scepsis, doubt): The participant does not believe something, is e.g. sceptical whether an idea is good, a statement is true, a solution will work.
Dominant (challenging): The participant shows dominant behaviour with respect to someone else, e.g. the participant is commanding, controlling or persuading others.
Defensive (apologetic): The participant reacts defensively to e.g. protect own ideas, or authority.
Supportive (affirmative, agreeing, approval): The participant shows support for another participant, either with respect to a contributions to the meeting, or towards his or her presence in the meeting.

Discussion

As we noted in the introduction. The various labels describe a mixed group of phenomena: emotions as well as meta-cognitive states and interpersonal stances. It is also clear that by providing a simple list the systematic ways in which some labels are connected does not come to the fore. For instance, one could think of the labels "curious" (attentive) and "distracted" (inattentive) to be opposites. However, this does not pose a problem as such. In this case, one could say that the format is equivalent to a seven point scale for the "attention" dimension (3 positive, 3 negative, and the neutral label). Interesting though, is the fact that some labels have opposites that are not in the list because they have not been marked in the corpus so far.
There is a group of labels that "confused", "surprised", "uncertain", "disbelief" that are not always easy to distinguish in context. Particularly because one thing may lead to another and cause and effect are, in general, hard to distinguish. Further trials should point out how confused annotators get applying these labels.

CONCLUSION

In this paper we have presented the annotation procedure for mental states of the AMI corpus. We are currently involved in extensive trials of the annotation schema. Results of reliability of the scheme, coverage, and the distribution of labels are presented in the main conference. The continuing trials should indicate the stability of the schema. Interesting to explore are the relations between these markers of mental states and other annotation levels, particularly the dialogue act schema and the argumentation scheme. The signals that annotators perceive that make them choose a particular label are also signals that the other participants in the meeting can see and will interpret as signs for the mental state. It will be interesting to see how these shape and determine the way the meetings proceed.

ACKNOWLEDGMENTS

We thank the annotators for all their efforts and comments. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication).

REFERENCES

1. Cowie, R. E. Douglas-Cowie, S. Savvidou, E. MacMahon, M. Sawey, M. Schröder (2000) 'FeelTrace': An instrument for recording perceived emotion in real time. ISCA Workshop on Speech and Emotion, p. 19-24
2. Reidsma, R. D. Heylen, R. Ordelman (2006) Annotating Emotion in meetings. Proceedings of LREC.

First Suggestions for an Emotion Annotation and Representation Language

Marc Schröder
DFKI GmbH
Saarbrücken, Germany
schroed@dfki.de

Hannes Pirker
OFAI
Vienna, Austria
hannes.pirker@ofai.at

Myriam Lamolle
LINC – University Paris 8 – IUT de Montreuil, Paris, France
lamolle@iut.univ-paris8.fr

ABSTRACT

This paper suggests a syntax for an XML-based language for representing and annotating emotions in technological contexts. In contrast to existing markup languages, where emotion is often represented in an ad-hoc way as part of a specific language, we propose a language aiming to be usable in a wide range of use cases, including corpus annotation as well as systems capable of recognising or generating emotions. We describe the scientific basis of our choice of emotion representations and the use case analysis through which we have determined the required expressive power of the language. We illustrate core properties of the proposed language using examples from corpus annotation.

INTRODUCTION

Representing emotional states in technological environments is necessarily based on some representation format. Ideally, such an Emotion Annotation and Representation Language (EARL) should be standardised to allow for data exchange, re-use of resources, and to enable system components to work together smoothly.

As there is no agreed model of emotion, creating such a unified representation format is difficult. In addition, the requirements coming from different use cases vary considerably. In the Network of Excellence HUMAINE, we have nevertheless formulated a first suggestion, leaving much freedom to the user to "plug in" their preferred emotion representation. The possibility to map one representation to another will make the format usable in heterogeneous environments where no single emotion representation can be used.

DIFFERENT DESCRIPTIVE SCHEMES FOR EMOTIONS

A unified theory or model of emotional states currently does not exist [1]. Out of the range of existing types of descriptions, we focus on three that may be relevant when annotating corpora, or that may be used in different components of an emotion-oriented technological system.

Categorical representations are the simplest and most wide-spread, using a word to describe an emotional state. Such category sets have been proposed on different grounds, including evolutionarily basic emotion categories [2]; most frequent everyday emotions [3]; application-specific emotion sets [4]; or categories describing other affective states, such as moods or interpersonal stances [5].

Dimensional descriptions capture essential properties of emotional states, such as arousal (active/passive) and valence (negative/positive) [6]. Emotion dimensions can be used to describe general emotional tendencies, including low-intensity emotions.

Appraisal representations [7] characterise emotional states in terms of the detailed evaluations of eliciting conditions, such as their familiarity, intrinsic pleasantness, or relevance to one's goals. Such detail can be used to characterise the cause or object of an emotion as it arises from the context, or to predict emotions in AI systems [8, 9].

USE CASES AND REQUIREMENTS FOR AN EMOTION ANNOTATION AND REPRESENTATION LANGUAGE

In order to ensure that the expressive power of the representation language will make it suitable for a broad range of future applications, the design process for EARL was initiated by performing a collection of use cases among members of HUMAINE. This list of use cases for emotional representations comprises i) manual annotation of emotional content of (multimodal) databases, ii) affect recognition systems and iii) affective generation systems such as speech synthesizers or embodied conversational agents (ECAs). On the basis of these use cases and the survey of theoretical models of emotions, a first list of requirements for EARL was compiled, which subsequently underwent discussion and refinement by a considerable number of HUMAINE participants.

Among the different use cases, the annotation of databases poses the most refined and extended list of requirements, which also covers the requirements raised in systems for recognition or generation.
In the simplest case, text is marked up with categorical labels only. More complex use cases comprise time-varying encoding of emotion dimensions [6], independent annotation of multiple modalities, or the specification of relations between emotions occurring simultaneously (e.g. blending, masking) [3].
EARL is thus requested to provide means for encoding the following types of information.
Emotion descriptor. No single set of labels can be prescribed, because there is no agreement – neither in theory nor in application systems – on the types of emotion descriptors to use, and even less on the exact labels that should be used. EARL has to provide means for using different sets of categorical labels as well as emotion dimensions and appraisal-based descriptors of emotion.
Intensity of an emotion, to be expressed in terms of numeric values or discrete labels.
Regulation type, which encodes a person's attempt to regulate the expression of her emotions (e.g., simulate, hide, amplify).
Scope of an emotion label, which should be definable by linking it to a time span, a media object, a bit of text, a certain modality etc.
Combination of multiple emotions appearing simultaneously. Both the co-occurrence of emotions as well as the type of *relation* between these emotions (e.g. dominant vs. secondary emotion, masking, blending) should be specified.
Labeller confidence expresses the labeller's degree of confidence with the emotion label provided.
In addition to these information types included in the list of requirements, a number of additional items were discussed. Roughly these can be grouped into information about the person (i.e. demographic data but also personality traits), the social environment (e.g., social register, intended audience), communicative goals, and physical environment (e.g. constraints on movements due to physical restrictions). Though the general usefulness of many of these information types is undisputed, they are intentionally not part of the currently proposed EARL specification. If needed, they have to be specified in domain-specific coding schemes that embed EARL. It was decided to draw the line rather strictly and concentrate on the encoding of emotions in the first place, in order to ensure a small but workable representation core to start with. The main rational to justify this restrictive approach was to first provide a simple language for encoding emotional states proper, and to leave out the factors that may have led to the actual expression of this state. Thus, EARL only encodes the fact that a person is, e.g., trying to hide certain feelings, but not the fact that this is due to a specific reason such as social context. Clearly, more discussion is needed to refine the limits of what should be part of EARL.

PROPOSED REALISATION IN XML

We propose an extendable, XML-based language to annotate and represent emotions, which can easily be integrated into other markup languages, which allows for the mapping between different emotion representations, and which can easily be adapted to specific applications.
Our proposal shares certain properties with existing languages such as APML [10], RRL [8], and EmoTV coding scheme [3], but was re-designed from scratch to account for the requirements compiled from theory and use cases. We used XML Schema Definition (XSD) to specify the EARL grammar, which allows us to define abstract datatypes and extend or restrict these to specify a particular set of emotion categories, dimensions or appraisals.
The following sections will present some core features of the proposed language, using illustrations from various types of data annotation.

Simple emotions

In EARL, emotion tags can be simple or complex. A simple <emotion> uses attributes to specify the category, dimensions and/or appraisals of one emotional state. Emotion tags can enclose text, link to other XML nodes, or specify a time span using start and end times to define their scope.
One design principle for EARL was that simple cases should look simple. For example, annotating text with a simple "pleasure" emotion results in a simple structure:

```
<emotion category="pleasure">Hello!</emotion>
```

Annotating the facial expression in a picture file face12.jpg with the category "pleasure" is simply:

```
<emotion xlink:href="face12.jpg"
category="pleasure"/>
```

This "stand-off" annotation, using a reference attribute, can be used to refer to external files or to XML nodes in the same or a different annotation document in order to define the scope of the represented emotion.
In uni-modal or multi-modal clips, such as speech or video recordings, a start and end time can be used to determine the scope:

```
<emotion start="0.4" end="1.3"
category="pleasure"/>
```

Besides categories, it is also possible to describe a simple emotion using emotion dimensions or appraisals:

```
<emotion xlink:href="face12.jpg" arousal="-0.2"
  valence="0.5" power="0.2"/>

<emotion xlink:href="face12.jpg"

        suddenness="-0.8"
        intrinsic_pleasantness="0.7"
        goal_conduciveness="0.3"
        relevance_self_concerns="0.7"/>
```

EARL is designed to give users full control over the sets of categories, dimensions and/or appraisals to be used in a specific application or annotation context (see below).

Information can be added to describe various additional properties of the emotion: an emotion *intensity*; a *confidence* value, which can be used to reflect the (human or machine) labeller's confidence in the emotion annotation; a *regulation* type, to indicate an attempt to suppress, amplify, or simulate the expression of an emotion; and a *modality*, if the annotation is to be restricted to one modality.

For example, an annotation of a face showing simulated pleasure of high intensity:

```
<emotion xlink:href="face12.jpg"
category="pleasure"
  regulation="simulate" intensity="0.9"/>
```

In order to clarify that it is the face modality in which a pleasure emotion is detected with moderate confidence, we can write:

```
<emotion xlink:href="face12.jpg"
category="pleasure"
  modality="face" confidence="0.5"/>
```

In combination, these attributes allow for a detailed description of individual emotions that do not vary in time.

Complex emotions

A <complex-emotion> describes one state composed of several aspects, for example because two emotions co-occur, or because of a regulation attempt, where one emotion is masked by the simulation of another one.

For example, to express that an expression could be either pleasure or friendliness, one could annotate:

```
<complex-emotion xlink:href="face12.jpg">
  <emotion category="pleasure"
confidence="0.5"/>
  <emotion category="friendliness"
confidence="0.5"/>
</complex-emotion>
```

The co-occurrence of a major emotion of "pleasure" with a minor emotion of "worry" can be represented as follows.

```
<complex-emotion xlink:href="face12.jpg">
  <emotion category="pleasure" intensity="0.7"/>
  <emotion category="worry" intensity="0.5"/>
</complex-emotion>
```

Simulated pleasure masking suppressed annoyance would be represented:

```
<complex-emotion xlink:href="face12.jpg">
  <emotion category="pleasure"
regulation="simulate"/>
  <emotion category="annoyance"
regulation="suppress"/>
</complex-emotion>
```

If different emotions are to be annotated for different modalities in a multi-modal clip, there are two choices. On the one hand, they can be described as different aspects of one complex emotion, and thus share the same scope, i.e. the same start and end time:

```
<complex-emotion start="0.4" end="1.3">
  <emotion category="pleasure" modality="face"/>
  <emotion category="worry" modality="voice"/>
</complex-emotion>
```

Alternatively, the expressions in the different modalities can be described as separate events, each with their own temporal scope:

```
<emotion start="0" end="1.9" category="pleasure"
  modality="face"/>
<emotion start="0.4" end="1.3" category="worry"
  modality="voice"/>
```

It is an open question which of these alternatives is most useful in practice.

Annotating time-varying signals

Two modes are previewed for describing emotions that vary over time. They correspond to types of annotation tools used for labelling emotional database. The Anvil [11] approach consists in assigning a (possibly complex) label to a time span in which a property is conceptualised as constant. This can be described with the start and end attributes presented above.

The Feeltrace [6] approach consists in tracing a small number of dimensions continuously over time. In EARL, we propose to specify such time-varying attributes using embedded <samples> tags.

For example, a curve annotated with Feeltrace describing a shift from a neutral state to an active negative state would be realised using two <samples> elements, one for each dimension:

```
<emotion start="2" end="2.7">
  <samples value="arousal" rate="10">
    0 .1 .25 .4 .55 .6 .65 .66
  </samples>
  <samples value="valence" rate="10">
    0 -.1 -.2 -.25 -.3 -.4 -.4 -.45
  </samples>
</emotion>
```

The output of more recent descendents of Feeltrace, which can be used to annotate various regulations or appraisals, can be represented in the same way. A sudden drop in the appraisal "consonant with expectation" can be described:

```
<emotion start="2" end="2.7">
  <samples     value="consonant_with_expectation"
rate="10">
    .9 .9 .7 .4 .1 -.3 -.7 -.75
  </samples>
</emotion>
```

This relatively simple set of XML elements addresses many of the collected requirements.

A FAMILY OF EARL DIALECTS: XML SCHEMA DESIGN

Our suggested solution to the dilemma that no agreed emotion representation exists is to clearly separate the definition of an EARL document's structure from the concrete emotion labels allowed, in a modular design. Each concrete EARL dialect is defined by combining a base XML schema, which defines the structure, and three XML schema "plugins", containing the definitions for the sets of emotion categories, dimensions and appraisal tags, respectively. Different alternatives for each of these plugins exist, defining different sets of category labels, dimensions and appraisals.

For example, to allow emotions to be described by a core set of 27 categories describing everyday emotions in combination with two emotion dimensions, the EARL dialect would combine the base schema with the corresponding plugins for the 27 categories and the two dimensions, and the "empty set" plugin for appraisals. Another EARL dialect, describing emotions in terms of four application-specific categories, would combine the base schema with an application-specific category plugin and two "empty set" plugins for dimensions and appraisals.

Even though EARL will provide users with the freedom to define their own emotion descriptor plugins, a default set of categories, dimensions and appraisals will be proposed, which can be used if there are no strong reasons for doing otherwise.

MAPPING EMOTION REPRESENTATIONS

The reason why EARL previews the use of different emotion representations is that no preferred representation has yet emerged for all types of use. Instead, the most profitable representation to use depends on the application. Still, it may be necessary to convert between different emotion representations, e.g. to enable components in a multi-modal generation system to work together even though they use different emotion representations [8].

For that reason, EARL will be complemented with a mechanism for mapping between emotion representations. From a scientific point of view, it will not always be possible to define such mappings. For example, the mapping between categories and dimensions will only work in one direction. Emotion categories, understood as short labels for complex states, can be located on emotion dimensions representing core properties; but a position in emotion dimension space is ambiguous with respect to many of the specific properties of emotion categories, and can thus only be mapped to generic super-categories. Guidelines for defining scientifically meaningful mappings will be provided.

OUTLOOK

We have presented the expressive power of the EARL specification as it is currently conceived. Some specifications are still suboptimal, such as the representation of the start and end times, or the fact that regulation types cannot be associated a numerical degree (e.g., degree of simulation). Other aspects may be missing but will be required by users, such as the annotation of the object of an emotion or the situational context. The current design choices can be questioned, e.g. more clarity could be gained by replacing the current flat list of attributes for categories, dimensions and appraisals with a substructure of elements. On the other hand, this would increase the annotation overhead, especially for simple annotations, which in practice may be the most frequently used. An iterative procedure of comment and improvement is needed before this language is likely to stabilise into a form suitable for a broad range of applications.

The suggestions outlined in this paper have been elaborated in a detailed specification, currently submitted for comment within HUMAINE. Release of a first public draft is previewed for June 2006. We are investigating opportunities for promoting the standardisation of the EARL as a recommended representation format for emotional states in technological applications.

ACKNOWLEDGMENTS

We gratefully acknowledge the numerous constructive comments we received from HUMAINE participants. Without them, this work would not have been possible.

This research was supported by the EU Network of Excellence HUMAINE (IST 507422) and by the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). OFAI is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology.

This publication reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

REFERENCES

1. Scherer, K. et al., 2005. Proposal for exemplars and work towards them: Theory of emotions. HUMAINE deliverable D3e, http://emotion-research.net/deliverables
2. Ekman, P. (1999). Basic emotions. In Tim Dalgleish and Mick J. Power (Ed.), *Handbook of Cognition & Emotion* (pp. 301–320). New York: John Wiley.
3. Douglas-Cowie, E., L. Devillers, J-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox (2005). Multimodal Databases of Everyday Emotion: Facing up to Complexity. In *Proc. InterSpeech*, Lisbon, September 2005.
4. Steidl, S., Levit, M., Batliner, A., Nöth, E., & Niemann, H. (2005). "Of all things the measure is man" - automatic classification of emotions and inter-labeler consistency. *ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing, March 19-23, 2005, Philadelphia, U.S.A., Proceedings* (pp. 317--320).
5. Scherer, K.R. (2000). Psychological models of emotion. In J. C. Borod (Ed.), *The Neuropsychology of Emotion* (pp. 137–162). New York: Oxford University Press.
6. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M.

(2000). 'FEELTRACE': An instrument for recording perceived emotion in real time, *ISCA Workshop on Speech and Emotion, Northern Ireland* , p. 19-24.

7. Ellsworth, P.C., & Scherer, K. (2003). Appraisal processes in emotion. In Davidson R.J. et al. (Ed.), *Handbook of Affective Sciences* (pp. 572-595). Oxford New York: Oxford University Press.

8. Krenn, B., Pirker, H., Grice, M., Piwek, P., Deemter, K.v., Schröder, M., Klesen, M., & Gstrein, E. (2002). Generation of multimodal dialogue for net environments. *Proceedings of Konvens*. Saarbrücken, Germany.

9. Aylett, R.S. (2004) Agents and affect: why embodied agents need affective systems Invited paper, 3rd Hellenic Conference on AI, Samos, May 2004 Springer Verlag LNAI 3025 pp496-504

10. de Carolis, B., C. Pelachaud, I. Poggi, M. Steedman (2004).APML, a Mark-up Language for Believable Behavior Generation, in H. Prendinger, Ed, *Life-like Characters. Tools, Affective Functions and Applications*, Springer.

11. Kipp, M. (2004). Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. Boca Raton, Florida: Dissertation.com