Proc. of LREC 2006, Genoa, p. XXV--XXVIII

Resources for the Processing of Affect in Interactions

Nick Campbell, Laurence Devillers, Ellen Douglas-Cowie,
Veronique Auberge, Anton Batliner, and Jianhua Tao

Abstract

Within the speech and language processing communitites there is considerable and growing interest in issues related to emotion and affect in speech (see e.g., the Humaine workshop held immediately prior to and as a satellite of this conference). However, the terms ”emotion” and ”affect” are often used almost interchangeably. The goal of this panel discussion will be first to define and differentiate the two terms, as they relate to speech processing, and then to specify the different needs and requirements of research and technology development for each. All panelists have experience in these fields of speech and language processing, and will be able to call on their own experience as well as that gained from discussions in the half-day workshop preceeding LREC ”Corpora for Research on Emotion And Affect”. We look forward to lively contributions from the floor and hope that the discussion will allow us to establish a common ground between the various disciplines engaged in collecting related corpora so that a better understanding of the needs of each community may be achieved.

1. Introduction

The guests of this panel session on “Affect in Interaction” were brought together for their experience in collecting and analysing large corpora of emotional or affect-marked speech and video data for use in speech and language technology. The success of e.g., the recent international conference on “Affective Computing and Intelligent Interaction” [1], and the developments from the European Network of Excellence Humaine [2], show that “emotion” and “affect” are beginning to be considered significant factors in the design of speech and language processing systems, perhaps even the next breakthrough on the way towards people-friendly systems that will be accepted for general use by non-experts and that will find their way into the homes and lives of ordinary people.

The aim of the discussion is to present an overview of the current state of the art with respect to the production and analysis of fundamental resources for the processing of affect and emotion-related information specifically from the point of view of research in speech and language technology. A primary goal of this session is to attempt to define the two terms ‘Affect’ and ‘Emotion’ in connection with their relevance to human speech communication, and to establish a common ground between the various disciplines engaged in collecting related corpora so that a better understanding of the needs of each community may be achieved.

It seems that “affect & emotion” is often used as a ‘bucket’ term that implies the two subcomponents to be synonymous or interchangeable . . . thus, as usual, the biggest problems seem to be related to terminology. We all use the same terms in our writing and in our thinking, but many of us use them to mean different things; and some of us even different things at different times.

2. Terminology

Anton Batliner comments: In the language and gender discussions there is a well-known phenomenon called “parasitic reference”:

> Tissues are called Kleenex; petroleum jelly, Vaseline; bleach, Chlorox, etc. to the economic benefit of the specific brands referred to and to the detriment of those brands that are ignored by this terminology. The alleged gender-neutral uses of “he”, “man”, etc. are just further examples. A gender-specific term, one that refers to a high-status subset of the whole class, is used in place of a neutral generic term. [3]

Thus people talked about emotions and intonation making themselves and others believe that this is the whole story. Of course, in fact, most people know that this is not exactly the case, and therefore, you quite often resort to some *rhetorical modification*: In the language and gender business: “In this paper, we are using the male form “he” but want to stress that we always talk about females and men alike.” In the emotion business: “In this paper, we use the term “emotion” in a very broad sense, not confined to the big-six, full-blown emotions. etc., etc.” In the intonation business: “We are using the term “intonation” in a broad sense ....” (although this last example is rather obsolete nowadays, almost all people talk about “prosody” and only about “intonation” in a narrow, restricted sense.) The crucial factor has of course been the concentration on elicited and acted data - and maybe the grounding in physiological, psychological theories.

Of course, we can continue to talk about emotions together with such rhethorical modifications. If we (are willing to) learn our lessons from the language and gender business, however: there are two possible strategies, first, making language more general, and second, making differences more visible.

3. Complexity of the Affective States in Real-Life Interactions

Laurence Devillers adds: In the computer science community, the widely used terms of “emotion” or “emotional state” are used without distinction from the more generic term “affective state”, which may be viewed as more adequate to describe the complex emotional state of a person. This “affective state” includes the emotions / feelings / attitudes / moods / and the interpersonal stances of a person. There is a significant gap between the affective states observed with artificial data (acted data or induced data) and those observed with real-life spontaneous data. This difference is mainly due to the context.

I define “context” here as the events that are at the origin of the affective state of a person, and these could be external or internal events. As an example of different events that can trigger different emotions / attitudes / interpersonal stances at the same time, we can imagine a physical internal event such as “a stomach-ache” that triggers pain and an external event as “someone helping the sick person” that triggers relief. In the artificial data, this context is “rubbed out” so we can expect to have much more simple full-blown affect states which are quite far away from real affective states.

If our goal is to build emotionally “non-caricatural” Human-Machine Interaction system, we have to focus on real-life databases and natural interaction with genuine emotional cause events instead of on biased data with artificial events.

The affective state of a person at any given time is a mixture of emotion/attitude/mood/interpersonal stance with often multi-trigger events occurring at different times. Past and recent events are often mixed to produce an affective state. At any given moment the brain is the seat of many emotions/affects, possibly with different valences. A person can feel relief because someone helps him/her but at the same time be sad. A politician may display positive attitudes masking his/her real disappointment after obtaining unexpectedly bad results in an election. Furthermore, the affective states are dynamic and constantly in change during an interaction.

4. Cognitive Aspects of Affect & Emotion

Veronique Auberge: Affect in speech is expressed following different cognitive processing levels, from involuntary controlled expressions – the so called emotions – to the intentional control of other kinds of affect that reveals the speakers intentions, attitudes, and linguistic expressivity (i.e., the choice of lexicon and grammatical para-phrases). The attitudes expressions (like “authority”, “doubt”, “surprise”, “politeness”, etc.) are socially and language dependent. They are the main part of face-to-face language interaction between humans in common life. Emotions are expressed only when the arousal vs. the inhibition are relevant, depending on the emotional context. They are rarely expressed in everyday life (see the Crest ESP corpus [4]), but the are decisive information for understanding.

To build an authentic expressive corpus in real life contexts can be done by selecting situations recorded in real life in function (see Cowie et al or Campbell) of the needed affects or by provoking real-life situations with expected affects (for example with a wizard of Oz paradigm, e.g. Auberge’ et al (2004)). One major problem is to label the affects expressed in the corpus.

In annotation tasks of linguistic or phonetic features, it has been shown that the confidence of the expert (a linguist or a phonetician) must be verified (see for example the SAM European project report). As for the social affect, since the affective features are described by conventional pragmatic labels, the problem is similar to linguistic labelling, and the expert can be trained to a scientific model developed for these social affects.

If a human can become an expert in emotion annotations (i.e., involuntary controlled affects), it is not because he has learned an objective scientific cognitive processing devoted to labelling, but because this expert uses the normal “naïve” human competences implied to consciously decode emotions in the ecological meta-situation built by a human (the expert) observing some others humans (the corpus) in a basic ecological situation.

That implies that the decoding has a limit of identification given by the empathy processing. This could also imply some artefacts, but it can be quite controlled by cumulating the experts, by verifying their coherence, and by completing this labelling by perceptive experiments.

5. How to Collect a Real-Life Corpus?

Laurence Devillers: Studying emotion raises several questions concerning ethics, naturalism of the emotion, contextual dependencies, etc. Recording actors may be aimed at providing controlled answers to these questions, but we should ask how to collect more real-life corpora.

We have conducted several experiments of “emotion” annotation with audio-only and audio-visual data extracted from real interviews or recorded in “call centers” which highlight the complexity of real-life emotional behavior.

Call centers provide interesting opportunities for recording people in various natural spoken emotional states, since the recordings can be made imperceptively and they provide real and genuine contexts where emotions are often exacerbated. For audio-visual data, it is necessary to find a context where the video-taped person can forget the presence of the camera and does not act. TV interviews during news, for example, are generally more natural than talk-shows or ‘reality’ TV.

Furthermore, the required size of such a corpus and the granularity of annotations depend on the research goals. For detection purposes, statistical approaches are greedy for large data sets. For realistic generation purposes, a fine-grained annotation of a smaller corpus might be more relevant.

5.1. Call-center data

We are looking at data recorded in a financial call center and in a medical call center. Our use of this data carefully respects ethical conventions and agreements ensuring the anonymity of the callers, the privacy of personal information and the non-diffusion of the corpus and annotations.

In certain states of mind, it is possible to exhibit more than one emotion; for instance, when trying to mask a feeling about something, when suffering, or when there are conflicting intentions, etc. We have found many manifestation of naturally-occurring mixed emotions in telephone dialogs recorded in the two call centers. For the corpus recorded at the financial call center, mixed emotions were observed for the clients combining Fear and Anger (or more appropriately *anxiety* and *annoyance*). Many clients show annoyance when they are fearful of losing money. This emotion mixture is never seen in the agents’ data.

In the second corpus, comprised of dialogs recorded in a Medical Help call center, specific emotion mixtures were found in different parts of the dialog: Agents showed impatience/anxiety mixtures when they identified a high level of emergency and experienced difficulties in dialoguing with the caller (e.g., difficulty in understanding non-native persons, social differences, physical condition, etc). 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/her 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.

5.2. Audio visual data

Audiovisual data are clips extracted from TV news with also a high degree of naturalness. our "EmoTV" video clips have been selected with the following constraints: Interviews during news (2 people, only one visible), no spoken feedback from the journalist who interviews, people are recorded in the same position in front of the camera with their upper body visible. Our corpus is well balanced between positive and negative emotions. It is also rich in blended emotions such as conflictual valences, i.e. positive relief blended with sadness. Clips also show rich emotional behaviors expressed by gestures, facial expressions and speech (prosody and verbal content). A lot of blended emotions show conflictual multimodal cues, by example, cry to bring relief.

5.3. Wizard-of-Oz

Veronique Auberge: We have been testing another alternative approach for affect annotation using a wizard of Oz corpus which consists in using the self consciousness of the subject who has felt the affects in the recorded situation. Our auto-annotation method has the advantage of being the closest to the subject, very precise for complex mixed affects and mental states expressions (e.g., "feeling of thinking"), but the naïve subject must be free to auto-annotate his own corpus without any constraints on the way to annotate it, and this is dependent on language since it has been shown that naive subject mainly rely on language to make explicit the emotions he felt.

"E-Wiz" (our Emotional Wizard of Oz) is a user-friendly freeware platform, developed at ICP as part of the JST/Crest Expressive Speech Processing project, in order to collect authentic but controlled, emotional, verbal, and non verbal interactions. The Sound Teacher E-Wiz scenario is presented as software to enable the subject to improve his or her phonetic mastery in the learning of languages. Our subjects have been selected to be strongly motivated by this task. The corpus (17 subjects, 15 hours) is multimodal (visual signal, speech signal, articulatory signal and bio-physiological signals). The auto-annotation has been partially verified with perceptive tests. The main part of the corpus is, unexpectively, the non speech part, which is very rich in complex annotations, mixing emotions, attitudes and mental states.

6. To Develop a Reliable Annotation Scheme

Laurence Devillers: One of the main challenges we address is the categorisation and annotation of real-life emotions, requiring the definition of a pertinent and limited set of labels and dimensions, as well as an appropriate annotation scheme. Furthermore, inter-labeler agreement and annotation label confidences are important issues to address.

In order to describe emotion, four main problems have to be dealt with: the dynamic aspect of emotions, the possible mixture of emotions, context-dependency, and the highly person-dependent nature of emotion expression. First, the dynamic aspect of emotions can be expressed as a continuous mark in an N-dimensional space or at a coarse level by a sequence of emotionally quasi-stable segments labeled with discrete verbal labels. Second, the mixture of emotions can be described using N-label categories with operators on them (blended, sequential, masking, ambiguous, etc.) or as a continuous mark in the complex emotion space. Third, some of the context and speaker-dependencies can be annotated as meta-data [5].

6.1. The description of emotion in everyday data

Ellen Douglas-Cowie: Our approach to emotion labelling draws on work by various teams (both practical and theoretical) over the past two years, and reflects the particular demands of working with the everyday emotional data that HUMAINE prioritises. The approach involves three stages:
Stage 1: Global emotion labelling
Stage 2: Trace labelling
Stage 3: Quantal labelling
The process starts with a broad emotion labelling applied at a global level — that is, emotion labels will be assigned across the whole emotional 'clip' or passage selected. The second stage gives finer time resolution for labels that stage 1 indicates are worth pursuing. The third stage, "quantal labelling" is the now standard kind of labelling where the timeline shows presence or absence of some attribute (anger, pointing, speech, . . . etc.).

There are three main reasons why quantal labelling might be useful. First, it has the potential to give finer and more accurate time resolution than trace labelling if (for instance) one wants to know the temporal relationship between a change in voice pitch and a change in emotion. Second, it offers the kind of qualitative representation that ECAs and speech recognisers have traditionally used, and we don't know whether you can handle the continuous information that the Trace programs provide. Third, tracing is not a very natural way to deal with some variables (such as whether an episode is emotionally pure, mixed, or moves through a sequence of related emotions.

The labels to be assigned fall under a number of headings:

Everyday emotion words: Labellers will select labels from a list of everyday emotion terms that best describe the emotion in the clip. Labellers are asked to select up to 6 labels from this list and to number them in order of best fit.

Types of emotion-related state Labellers will identify the types of emotion-related state that are present using labels from a list that we have developed at QUB and validated experimentally.

Combination types Labellers will select labels from a list (due to LIMSI) to describe the emotion combinations that occur in the clip: unmixed simultaneous combination (distinct emotions present at the same time) sequential combination (single episode involving a sequence of related emotions)

Authenticity Labellers will specify whether the clip appears to involve misdirection about the person's actual emotions. The labels will involve two scales Acted (from no acting of emotion to extreme acting of emotion) Masked (from no concealment of emotion to total concealment of emotion) Two labels will be given for each (to mark the range of states involved, from most to least authentic).

Core affect dimensions Labellers will specify the extremes of core affect observed in the clip, i.e. maximum observed intensity, maximum observed activation, most extreme positive valence observed, most extreme negative valence observed. The labelling will use standard scales for intensity, activation, and valence.

Appraisal categories Labellers will specify how strongly the person's emotional state is related to selected appraisal factors. The factors are Goal conduciveness (the situation offers an opportunity to achieve a significant goal); Goal obstructiveness (the situation blocks the person from achieving a significant goal); Power / powerlessness (how the person rates his/her ability to affect events). These are selected on the basis of previous work which indicates that they can be rated reasonably reliably in naturalistic material.

6.2. Validation Protocols

Laurence Devillers: After the language and level of representation decisions (context annotation at the dialog level, emotion annotation at the segment level, label definition) are made, an annotation and validation protocol needs to be defined, assuring that the reference 'emotions' can be accurately extracted for use with machine learning.

The high subjectivity of human annotation requires the use of rigorous annotation protocols. The emotional units can be at the level of the speaker turn, the segment, or the word. The segments (within a speaker turn) can be defined as a syntactic or semantic group. Concerning label consistency, it is evident that combining the opinions from a larger number of annotators (at least 3) via majority voting, for example, leads to less subjective annotations. Evidently, the larger the size of the corpus, the more difficult it is to obtain multiple annotations. We also have to consider inter-labeler consistency and confidence measures.

There are different measures of annotation reliability, and the perception of emotion is very subjective, for instance, some persons are more compassionate (or receptive) than others. How does this affect the reliability of the annotations? Can annotators be wrong in their perceptions? Are there good and bad annotators? In our opinion, a good labeler is coherent over time and is able to explain his/her decisions. Just as the expression of emotion is highly personal, so is its perception. Our philosophy is to exploit these differences by combining the labels from multiple annotators in a soft emotion vector. How to then use this vector effectively in machine learning is one of our future objectives.

Ellen Douglas-Cowie: Labellers should also remember to address the following broad context classes: Communicative context, Communicative goals, Target audience, Camera awareness, Physical context, etc.

7. Summary

Jianhua Tao and Nick Campbell: It is clear from the above discussion that this field is making some very rapid progress and that we are now beginning to see a consensus both in the collection of data and in the terms used to describe it. We see too, that far from being as simple as "emotion" in speech and video, it is a subtle blend of many more complex and often seemingly contradictory factors that are very relevant to human communication and that are perceived without any conscious effort by any native speaker of the language or member of the same cultural group.

If machines are to be made sensitive to this type of information, which we believe to be as important to human communication as (or perhaps even more so than) the linguistic or propositional content, then we will need more such corpora upon which to base our research. If these corpora are acted or contrived, then the resulting technology will be of little use; the more natural the data we collect, and the more complex the factors they contain, the closer we can come to understanding the mechanisms of human social communication and perhaps modelling them, for use in the ubiquitous computing that is becoming so much a part of our lives.

8. Acknowledgements

The panel is chaired by Nick Campbell of the National Institute of Information & Communications Technology, Japan, and the Acoustics and Speech Processing Department of the Spoken Language Communication Research Labs at ATR in Japan, and by Jianhua Tao of the National Laboratory of Pattern Recognition (NLPR) in the Institute of Automation of the Chinese Academy of Sciences. The four guests are Laurence Devillers of CNRS-LIMSI in France, Ellen Douglas-Cowie of the Queen's University Belfast, Ireland, Veronique Auberge of the ICP-CNRS in France, and Anton Batliner of Erlangen University in Germany. We expect half of the discussion to come from the floor and will therefore, subject to the reviewers' approval, summarise the key topics of the discussion in a subsequent paper to be presented at the next LREC conference.

References

[1] International Conference of Affective Computing and Intelligent Interaction, http://www.affectivecomputing.org/

[2] The European Network of Excellence, Humaine: web pages at http://emotion-research.net/

[3] Moulton, J., "The myth of the neutral man", pp. 100-115 in Vetterling-Braggin, ed, 1981.

[4] The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.atr.jp

[5] Devillers, Cowie, Martin, Douglas-Cowie, et al. LREC 2006.