A Taxonomy of Applications that Utilize Emotional Awareness

Anton Batliner[†], Felix Burkhardt[*], Markus van Ballegooy[*], Elmar Nöth[†]

[†]Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg, Martensstr. 3,
91058 Erlangen, Germany
{batliner,noeth}@informatik.uni-erlangen.de
∗T-Systems Enterprise Services GmbH, Goslarer Ufer 35, 10589 Berlin, Germany
{Felix.Burkhardt,Markus.van-Ballegooy}@t-systems.com

Abstract

This paper deals with human-computer interaction applications that utilize emotional awareness. We will confine our discussion on speech-based applications. Prerequisites — training data, annotations — as well as state of the art in recognition and synthesis are addressed focusing on usability in possible applications and keeping restrictions in industrial environments in mind. We will present a taxonomy of applications using criteria such as online/offline, mirroring/non-mirroring, emotional/non-emotional and critical/non-critical system reactions. Based on a list of prototypical applications we check the consistency and usefulness of this taxonomy.

Taksonomija aplikacij, ki uporabljajo čustveno zavedanje

Prispevek se ukvarja z aplikacijami za interakcijo med človekom in računalnikom, ki uporabljajo čustveno zavedanje. Diskusijo omejujemo na govorne aplikacije. Obravnavamo predpogoje — učni podatki, označevanje — in najnovejša dognanja v razpoznavanju in sintezi govora, osredotočamo se na uporabnost možnih aplikacij ob upoštevanju omejitev v industrijskih okoljih. Predstavljamo taksonomijo aplikacij glede na to, ali omogočajo sprotni odziv ali ne, ali so zrcaljene ali ne, ali so čustvene ali ne ter ali so reakcije sistema kritične ali ne. Na podlagi seznama prototipskih aplikacij preverjamo doslednost in uporabnost te taksonomije.

1. Introduction

Taking into account emotional and emotion-related (affective) states of users interacting with automatic systems – be this automatic dialog systems or automatic systems in general – has been an emerging topic during the last years. The idea behind is simple: even if automatic systems were much better than state-of-the-art systems are today at interacting with human users, there is still a certain quality missing, namely recognizing and dealing with not only the semantics of the user's message but his/her emotional state as well. Often it is supposed that only then, the machine is really on an equal footing with the human communication partner.

The channels monitored and used by the system have probably been the most obvious way of telling apart different types of applications – alone because traditionally, they belong to different disciplines: speech (and by that, acoustic and/or linguistic information), facial expressions, gestures, body postures, and background knowledge (i.e. context in its broader sense). In this paper, we want to concentrate on systems that deal with speech only (recognition and generation/synthesis), and we want to introduce some further characteristics which can be used for a tentative taxonomy. By that, we concentrate mostly on choices made by the system designers and much less on the choices which are made – more or less consciously – by users of such systems.

In the history of emotion research, some prototypical application examples emerged, cf. e.g. (Picard, 1997). We introduce such a list of different application-sketches here and will use them as examples in the coming chapters of this article, representing substitutes or prototypes for similar applications that may be used in different fields.

Emotional Monitoring: E.g. anger detection can be used to soothe disgruntled users or for automatic quality monitoring.

Emotional Mirror: Speech analysis can be used for self-training (e.g. '*Do I sound boring?*')

Understanding Tutor: Teacher-student communication can be enhanced enormously by emotional channels in order to monitor and augment motivation.

Emotion-aware Surrounding: Quite an old idea is a computer controlled environment that adapts automatically on the user's mood by e.g. playing '*just the right music*' or adjusting automotive system reaction.

Believable Agent: The naturalness of an artificial '*being*' and the appearance of intelligence (we will not go into philosophical questions here) is highly altered by emotional expressions; especially gaming applications can benefit.

Emotional Chat: The high success of the so-called '*Emoticons*' shows how strong the human desire is to express emotion in mode-restricted computer mediated communication (CMC). Special channels can be provided to facilitate this and analysis can be used to automate emotional labeling.

This article is structured as follows: section 2. will discuss important aspects for data collection from the application point of view. The following Section 3. deals with technical potentialities of recognition and synthesis. Because applications will not be greenfield developments, we include a subchapter on industrial requirements. In Section 4. we propose our taxonomy and classify our example applications accordingly. We conclude with final remarks in Section 5.

2. Models need Data

Emotion-aware applications are based on models and models rely on data. Thus we will sketch some central issues with respect to data-collection from the application's perspective in this section. Although it is trivial to remark that the data should be as close to the intended application as possible (a self-training application would be perfect), it is of course not economic to collect data for each application. In order to reuse data-collections, standardized ways to annotate data are required and will emerge.

In contrast to other speaker characterizations like age or gender, emotion is a fuzzy topic. The performance of a human labeler can be measured by comparison with other labelers. Standardized ways to control performance of labelers, measurements of inter-labeler agreement (Steidl et al., 2005) and finding unified labels will be essential for application deployment. In many applications, e.g. the anger-detecting voice portal, the recognition of emotion and the system's reaction must play hand-in-hand, and dialog designer and labeler must rely on a common emotion-coding language.

Because emotional expression depends highly on speaker idiosyncrasies, speaker-dependent modeling should be preferred, if possible. Such personalized applications could require an emotion-recognition training process just like dictation systems do nowadays.

One of the problems that arises in data-collection and makes Heisenberg's uncertainty principle come into mind, comes from the fact that people, if they know they will be monitored, react differently then if they had not be aware of the monitoring; this phenomenon has been called 'observer's paradox' (Labov, 1970). On the other hand hidden monitoring is often difficult, probably unethical and generally prohibited by law.

3. State of the Art in Speech Technology

3.1. Emotion Recognition

The state of emotion recognition in general still suffers from the prevalence of acted laboratory speech as object of investigation. The high recognition rates of up to 100% reported for such corpora cannot be transferred onto realistic, spontaneous data. For realistic databases, performance for a two-class problem is typically $< 80\%$, for a four-class problem, $< 60\%$, cf. (Batliner et al., 2005); normally, acoustic (mostly prosodic and/or MFCC based features) and some plain linguistic features such as bag-of-words are employed. Performance can be improved

- by employing highly sophisticated classifiers,
- by concentrating on prototypical, clear cases ((Batliner et al., 2005) report up to 77.5% for a four-class problem),
- by mapping onto cover classes (for instance, only taking into account positive vs. negative valence),
- by taking into account cost functions (for instance, penalizing only 'severe' confusions),
- by resorting to speaker-dependent[1] modeling, as indicated above, if this is suitable for the resp. application,
- and by using other, additional knowledge bases (for instance, dialog and/or interaction history).

Larger databases, i.e. more training data, seem to be a must but are difficult to obtain because the reference (ground truth, i.e. the phenomena that have to be recognized) cannot be obtained easily: for word recognition, a simple transliteration will suffice; for emotion recognition, manual annotation is normally necessary, time-consuming and costly. In some few scenarios, it might be possible to resort to external evidence, allowing a sort of automatic annotation; for instance, in a car driving scenario, actual speed and movements of the steering wheel could be monitored automatically and used for labeling training data.

3.2. Approaches in Speech Synthesis

First attempts to simulate emotional speech by means of speech synthesis started soon after the first mature speech synthesizers were developed. For an overview on the history of emotional speech synthesis, the reader may be referred to (Schröder, 2001). Most of today's research concerning emotional speech synthesis is still dealing with a small set of basic emotions (e.g. the so-called 'big *n*', *n* being a small number like 4,5,6) such as anger, sadness, fear, or joy.

Because speech synthesis until now still has to solve more pressing problems than the simulation of emotional speech, and applications for emotional synthesis are yet more in the future than for emotion-recognition[2], the research is less advanced. The simulation of realistic and natural emotional speech expression is vital for most application scenarios, e.g. for applications to enhance the believability of talking heads. Some ideas of yet mainly unsolved problems that might be required by prospective applications are summarized in the following items:

- Simulation of a larger set of discrete emotions that are displayed more subtle than 'the big *n*' performed in a cartoon-like style.
- Blending between two or more emotions, like e.g. anger and sadness, and finding models for the transition from one emotion to a different one during one utterance.
- Finding acoustic correlates for other models than discrete emotions like emotional dimensions or stimulus evaluation checks in order to support different emotion models directly.
- There is still a big gap between system-modeling formant-synthesis (flexible but unnatural) and

[1] For speaker-dependent modeling, spectral features might come into play as well which might not be suitable for speaker-independent modeling.

[2] Consider that applications that utilize emotional synthesis often depend on artificial intelligence in order to know when to speak with which emotion.

manipulation-avoiding speech-concatenating non-uniform unit-selection (high quality but inflexible). Solving the problem of the discrepancy between naturalness and inflexibility of non-uniform unit selection approach is vital for emotion-simulation, a fact that becomes manifest e.g. in voice-quality modeling.

For the time being emotional high quality synthesis constrains in either adding some emotional interjections to the data, cf. (Eide et al., 2003), or reducing the set of emotions to a binary choice, e.g. agitated speaking style vs. normal. Note that this is easier in low-quality / small footprint approaches that include signal manipulation like formant- or diphone-synthesis.

3.3. Industrial Requirements

Emotion recognition in an industrial environment has to fulfill a set of requirements that should be considered while planning an emotion-aware application:

- The emotion aware modules must integrate into the existing architecture and should use standardized interfaces as much as possible.
- In an HCI system the delay caused by the processing must not obstruct the dialog flow.
- For most applications a classification task must be based solely on automatically gained features.
- The recognition algorithms have to work often on highly noisy data and performance results from laboratory studies are not directly applicable.
- The procedures must attend to economic issues, e.g. algorithms that are IPR (intellectual property rights) protected must be avoided and manual labor should be restricted.

With a growing dispersion of emotional applications a market of emotion-aware components operating on different processing steps will emerge; data vendors will collect emotion-annotated data to train the models, speech core-technology vendors will offer emotion recognition and simulation components, integrators will offer 'emotion-modules' for dialog platforms and gaming engines. This market will rely on a set of standards yet to be developed.

4. Applications

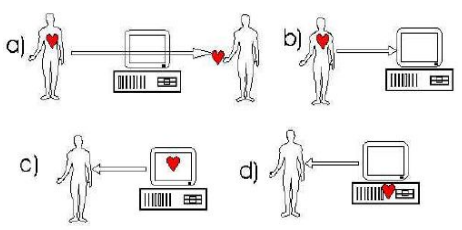

Figure 1: Different uses of emotional processing in computer systems.

Emotional applications can be thought up in an arbitrary number. This Section will give a frame to classify applications based on common features in order to facilitate the process of thinking up new useful application scenarios and identify reuse of common modules for existing ones.

Emotions can be processed in several places of an information-processing system (Picard, 1997). Figure 1 displays several possibilities:

a) Broadcast: Emotional expression is an important channel of information in human communication. In telecommunication it might be desirable to provide for a special channel for emotional communication. A popular example are the so-called '*emoticons*' used in e-mail communication.

b) Recognition: The human emotional expression can be analyzed in different modalities, and this knowledge is used to alter the system reaction.

c) Simulation: Emotional expression can be mimicked by the system in order to enhance a natural interface or to access further channels of communication, like e.g. uttering urgent messages in an agitated speech style.

d) Modeling: Internal models of emotional representations can be used to represent user- or system states or as models for artificial intelligence, e.g. influence decision making.

Of course systems may combine several of these features for an integrated application like the above mentioned emotional tutoring system. As we are focused on speech processing in this article, we will not follow up on the AI-possibilities but confine on the recognition, transport and simulation for communicative reasons.

Instead of identifying system components, we can use different features to tell apart emotional systems: In Table 1, we present four binary features telling apart different types of applications: +/- {*online, critical, mirroring, emotional*}. In principle, this would result in 16 possible combinations, but some combinations are less likely than others, e.g. an *offline* application cannot show *mirroring* as defined in the table and need not be *critical* as there can be a human check on the performance. In order to check the validity of our taxonomy we will classify our example applications accordingly. Most of these applications are already on the market or in a prototype stadium.

Emotional Monitoring: One of the most often expressed ideas for emotional speech monitoring are voice portals that use detection of negative feelings such as anger to appease by *mirroring* their expressions (Burkhardt et al., 2005). This application is *critical* because users could really get angry if they were – wrongly – 'accused' being angry but are not. It is an *online* application because the system reacts directly but could be imagined in an *offline* scenario, where the customer satisfaction is measured later by classifying the call logs. Depending on the system reaction it might be *emotional*, i.e. try to soothe the user

features	description
	system design
online	system reacts (immediately/delayed) while interacting with user
offline	no system reaction, or delayed reaction after actual interaction
mirroring	user gets feedback as for his/her emotional expression
non-mirroring	system does not give any explicit feedback
emotional	system reacts itself in an emotional way
non-emotional	system does not behave emotionally but 'neutral'
	meta-assessment
critical	application's aims are impaired if emotion is processed erroneously
non-critical	erroneous emotion processing does not impair application's aims

Table 1: Criteria for Taxonomy

by adequate dialog strategies; in a *non-emotional* variant, it can simply transfer to a human agent. A *non-mirroring* variety might result in ethical concerns: it could be used to monitor call center agents or psychotic patients and enable supervisors intervene if necessary. Another *non-mirroring* variety consists of automatically identifying untrustworthy customers, e.g. in an insurance portal.

Emotional Mirror: The emotional mirror was often suggested by R. Picard's team and was developed in the Jerk-o-meter application (Madan et al., 2005). A person's speech is monitored for emotional expression which can be used for training reasons. This application is definitely *mirroring* and could be used *online* as well as *offline*. It is *critical* because emotion recognition is the central aspect of the application which will, however, not react itself *emotional*. Note that for this application, training data could perhaps be acted because the user's intention might be to sort of act him/herself.

Understanding Tutor: Automatic tutoring is an interesting topic in a growing information society and emotional strategies are important to enhance motivation. Several systems have been suggested already, e.g. (Poel et al., 2004). This application must be *online* and *mirroring* to react directly, e.g. on the pupils boredom, and will be *emotional* to motivate the pupil but is not necessarily *critical*, i.e. reactions could be quite subtle.

Emotion-aware Surrounding: Quite an old idea is a computer controlled environment that adapts automatically on the user's mood by e.g. playing '*just the right music*' or adjusting automotive system reaction. This set of applications are of course *online* and *mirroring* but are not *emotional* themselves. Whether they are *critical* would be a distinction between an emotional CD-player and a car reacting on a stressed driver.

Believable Agent: The naturalness of artificial '*beings*' and the appearance of intelligence is highly altered by emotional expression; especially gaming applications can benefit. This application could be regarded as generic term for the understanding, emotional tutor and can consistently be classified just as this.

Emotional Chat: In an avatar-based chat system the avatar's emotional expression could be controlled by the user deliberately. An example where the emotional expression gets measured automatically would be Picard's classroom barometer application (Picard, 1997) where the tele-conference teacher is informed about the pupils attention via affective jewelry. It is an *online* (if automatically gained), *mirroring*, *non-emotional* application likely to be *non-critical* (if self-reported).

5. Concluding Remarks

In this article, we sketched necessary prerequisites for emotional systems such as data, recognition, and synthesis; in the resp. sections, we could not go into detail and only addressed some pivotal topics. Our main contribution is a tentative taxonomy of emotional applications. Taxonomy as such makes life easier: we know what we are looking for, and what we should decide between. It can be a sort of roadmap what features such as given in Table 1 to incorporate or to disregard for a new application, and to help thinking of new ways of incorporating emotional awareness. Such features or feature combination can characterize modules that can be turned on or off, depending on user characteristics, dialog step, and confidence measures.

So far, we have mostly talked about technology.[3] As long as emotional applications are in their infancy, this might seem OK. What has not been addressed is ethical issues – a topic which will be more pressing the better such systems perform. Emotion aware and expressive applications might seem more intelligent and capable than they really are. Thus it might not always be the best idea to make a system react emotionally, and it might not always be desirable because of ethical reason to make the user conceive of the system as human-like – not to speak of other ethical questions that will arise. Last but not least: just like convictions should not be based on lie-detectors, decisions that impact peoples' welfare severely should not be based on emotion recognition.

[3] Some more basic questions are dealt with in (Picard, 1997; Picard, 2003).

Acknowledgments: This work was partly funded by the EU in the framework HUMAINE (*http://emotion-research.net/*) under grant IST-2002-507422, and by the German Federal Ministry of Education and Research (*BMBF*) in the framework of SmartWeb (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

6. References

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

F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber. 2005. An emotion-aware voice portal. In *Proc. Electronic Speech Signal Processing ESSP*, pages 123–131.

E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli. 2003. A corpus-based approach to expressive speech synthesis. In *Proc. ISCA ITRW on Speech Synthesis, Pittsburgh*, pages 79–84.

W. Labov. 1970. The Study of Language in its Social Context. *Studium Generale*, 3:30–87.

A. Madan, R. Caneel, and A. Pentland. 2005. Voices of attraction. In *Proc. Augmented Cognition, HCI 2005, Las Vegas*.

R. Picard. 1997. *Affective computing*. MIT Press.

R. Picard. 2003. Affective Computing: Challenges. *Journal of Human-Computer Studies*, 59:55–64.

M. Poel, R. op den Akker, D. Heylen, and A. Nijholt. 2004. Emotion based agent architectures for tutoring systems: The ines architecture. In *Cybernetics and Systems 2004. Workshop on Affective Computational Entities (ACE 2004)*.

M. Schröder. 2001. Emotional speech synthesis - a review. In *Proc. Eurospeech 2001, Aalborg*, pages 561–564.

Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth, and Heinrich Niemann. 2005. "Of All Things the Measure is Man" - Classification of Emotions and Inter-Labeler Consistency . In *Proc. of ICASSP 2005*, pages 317–320.