Real-time vocal emotion recognition in artistic installations and interactive storytelling: Experiences and lessons learnt from CALLAS and IRIS

Thurid Vogt, Elisabeth André, Johannes Wagner
Lab for Multimedia Concepts and Applications
University of Augsburg, Germany
http://interactive-multimedia.de

Steve Gilroy, Fred Charles, Marc Cavazza
School of Computing
University of Teesside, Middlesbrough, UK
http://ive.scm.tees.ac.uk

Abstract

Most emotion recognition systems still rely exclusively on prototypical emotional vocal expressions that may be uniquely assigned to a particular class. In realistic applications, there is, however, no guarantee that emotions are expressed in a prototypical manner. In this paper, we report on challenges that arise when coping with non-prototypical emotions in the context of the CALLAS project and the IRIS network. CALLAS aims to develop interactive art installations that respond to the multimodal emotional input of performers and spectators in real-time. IRIS is concerned with the development of novel technologies for interactive storytelling. Both research initiatives represent an extreme case of non-prototypicality since neither the stimuli nor the emotional responses to stimuli may be considered as prototypical.

1. Introduction

It is well-known that speakers differ significantly in the expressivity of their voice. While it is hard to guess for some speakers in which emotional state they are, others reveal their emotional state quite clearly through their voice. The way they show emotions may be called prototypical, and independent observers would largely agree on the emotional state of these speakers. A common example includes voice data from actors for which developers of emotion recognition systems reported accuracy rates of over 80 % for seven emotion classes [11, 14, 16]. In realistic applications, there is, however, no guarantee that emotions are expressed in a prototypical manner. As a consequence, these applications still represent a great challenge for current emotion recognition systems.

When dealing with non-prototypical emotions, we have to analyze the causes of non-prototypicality. Non-prototypical behaviors may be elicited externally by certain events, or the behavior as such may be non-prototypical,

e. g. because some persons may in general not show their emotions clearly. That means the exclusion of non-prototypical stimuli when designing an application is not a guarantee that no prototypical behaviors will occur. Most recognition systems in the literature are based on machine learning methods. That is a large amount of emotional data is collected for which classifiers are trained and tested. To ensure satisfying recognition rates, it is decisive that the emotion-eliciting events during training are similar to the emotion-eliciting events during testing. Current systems either limit training on those prototypical cases for which a high interlabeler agreement could be achieved or train their system based on standard stimuli, for example emotions produced from professional actors. Both methods are, however, problematic in realistic applications.

In the following, we report on experience gained in the CALLAS[1] project and the IRIS[2] network where non-prototypicality was indeed identified as a major problem. CALLAS (**C**onveying **A**ffectiveness in **L**eading-edge **L**iving **A**daptive **S**ystems) is an Integrated Project funded by the EU which aims to develop interactive art installations that respond to the multimodal emotional input of performers and spectators in real-time. IRIS (**I**ntegrating **R**esearch in **I**nteractive **S**torytelling) is an EU-funded Network of Excellence concerned with the development of novel technologies for interactive storytelling. Recognition of affect is one of the novel techniques to be integrated into virtual storytelling environments. Affective input from the voice is analyzed in both research initiatives by our component EmoVoice [17] for speech emotion recognition which is described in more detail in Section 2.

The following showcases in CALLAS make use of EmoVoice to detect emotions from the user's voice or employ parts of EmoVoice to analyze acoustic features of emotional speech input:

- *E-Tree:* E-Tree by Teesside University [6] is an Aug-

---

[1] http://www.callas-newmedia.eu
[2] http://iris.scm.tees.ac.uk

Figure 1. E-Tree reacting to emotional input: negative/low-arousal, neutral and positive/high-arousal

mented Reality art installation of a virtual tree that grows, shrinks, changes colors, etc. by interpreting affective multimodal input from video, keywords and emotional voice tone (Fig. 1).

- *Galassie:* Galassie by Studio Azzurro[3] [5] creates stylized shapes similar to galaxies for each present user. The visual appearance of the galaxies depends on the user's emotional state which EmoVoice detects from the user's voice (Fig. 2).

- *PuppetWall:* In the PuppetWall showcase by Helsinki University of Technology (TKK) [9], a user may influence a 2D graphics by the movements of physical puppets and the emotional tone of his or her voice. In contrast to the other showcases where the system responds to emotional states, PuppetWall is controlled by acoustic features of the user's voice (Fig. 3).

- *Interactive Opera:* Interactive Opera by Digital Video[4] is a live performance for children recreating characters and sceneries of famous compositions from W.A. Mozart, Giacomo Puccini, Giuseppe Verdi and many others. The children may influence the outcome of the story by expressing emotions using facial expressions and their voice that are mapped onto the characters (Fig. 4).

- *Music Kiosk:* Music Kiosk by XIM[5] [10] is an interactive museum installation that presents music instruments to young people in an innovative way enabling them to control music by expressing their feelings (Fig. 5).

- *ElectroEmotion:* ElectroEmotion by TKK is an affective, interactive installation for public spaces that mainly served to collect a multimodal corpus of emotion data. In this showcase, users are directly requested to express different kinds of emotions which are visualized to provide the users with feedback.

[3]http://www.studioazzurro.com
[4]http://www.toonz.com
[5]http://www.xim.co.uk

Figure 2. Galassie

Figure 3. PuppetWall

Within IRIS, EmoEmma, an interactive storytelling system based on Gustave Flaubert's novel "Madame Bovary", has been developed by Teesside University. Users can influence the outcome of the story by acting as one of the characters and their interaction mode is restricted to the emotional tone of their voice (Fig. 6).

Most showcases are fully implemented and have been used already by real users, partly in user studies (EmoEmma, E-Tree, ElectroEmotion), but also under realistic conditions. For example, Galassie has been performed in July 2008 at Teatro Arcimboldi in Milan, Italy, and E-Tree has been presented to user at the EC's ICT (Information and Communication Technologies) event 2008. For

Figure 4. Interactive Opera

Figure 5. Music Kiosk

Figure 6. EmoEmma

this reason, our experiences are really gained under real-life conditions.

In all cases, EmoVoice was used to analyze the users' vocal emotions in real-time while they were interacting with the installation. Some showcases were completely controlled by the user's emotional state. That is, there was no analysis of the semantic content. In most cases, the user's emotional state was reflected by the system's display. For example, in the E-Tree installation there was a direct mapping between the user's emotional state and the color and size of the tree. In EmoEmma and Interactive Opera, the system did not simply mirror the user's emotional state. In-

stead a more sophisticated reasoning process was required to determine the output of an appropriate system response. For example, in EmoEmma, the user was able to influence by the emotional tone of his voice whether Emma would become unfaithful or not. Some showcases, such as the Music Kiosk and Interactive Opera, were specifically designed as multi-user applications.

Emotions expressed by the users of the showcases are very varied. They range from rather prototypical emotions in E-Tree, where users explore which emotional expressions can make the tree respond, to absolutely unprototypical and unpredictable expressions e. g. of spectators in Galassie. Showcases are intended to support artists or spectators to express themselves emotionally. Therefore, in most cases, allowed emotions cannot be imposed on the users. Furthermore, everybody has their own individual interpretation of emotions and expresses them differently. A number of further factors contribute to non-standard expressions of emotions in CALLAS: Some showcases have a limited, pre-selected user group, while Interactive Opera addresses children. For other showcases, in turn, potential users cannot be restricted. Languages comprise English, by native and non-native speakers, Italian, and Finnish. Background noises also vary a lot among and within showcases. E. g. E-Tree has been performed in quiet office environments and at exhibitions. In Galassie, Interactive Opera and Puppet-Wall, the background noise level is high and the voice of the speaker to be analyzed competes with other voices in the background. Though we recommend that every speaker wears a headset microphone, this is not feasible in some showcases. Artistic emotions often are rather exaggerated (but not necessarily prototypical at the same time), which might ease recognition, however, expectations of users are very high. In Galassie, users expected the system to be as sensitive as humans, or even more, to their emotional state and to understand even very subtle emotions. This is of course an expectation that cannot be met with current technology. But also the other demands go beyond those encountered at the offline analysis of existing emotional speech databases as done by previous and many current work on speech emotion recognition and require new strategies.

In the remainder of the paper we will first shortly describe EmoVoice, our speech emotion recognition component, on which our experiences of real-time emotion recognition are based, and present attempts by EmoVoice to deal with non-prototypical emotions in CALLAS and IRIS. Then we will discuss the problem of getting appropriate training data for classifiers suitable for the showcases. The fifth section concerns methods to evaluate emotion recognition in real-time artistic installations. Last, we will give some concluding remarks.

2. Real-time emotion recognition from voice: EmoVoice

EmoVoice [17] is a system for emotion recognition from voice and provides tools for acoustic feature extraction and building an emotion classifier as well as for recognizing emotions in real-time.

Acoustic features are derived from pitch, energy, voice quality, pauses, spectral and cepstral information as conveyed in the speech signal. In total, a set of 1451 acoustic features is calculated which can be reduced by standard feature selection methods. No semantic or word information is used, in order to make the recognition process faster, as no speech recognizer is necessary, and also when selecting feature extraction algorithms, attention was paid to speed. Integrated classifiers are Support Vector Machines and Naïve Bayes, while the latter one is used more often because it is faster and thus responds better to real-time demands.

In EmoVoice, classifier creation is supported by two user interfaces. The first interface allows recording a database of emotional speech, by reading a set of emotion inducing sentences or free speech input. With the second interface, a classifier can be trained, and a quality check of the classifier can be performed. Thus, personal or application-specific recognizers can be built without deep technical knowledge of the recognition process. The resulting classifier can be used by a command line tool that continuously classifies user speech (without push-to-talk) and that can be linked to any application by socket communication.

Speech segmentation is a critical aspect as it should be fast and at the same time provide meaningful and consistent segments. We found an automatic voice activity detection with no in-between pauses longer than 1 sec to be a good compromise between speed and accuracy. A similar strategy has been employed e. g. on the FAU Aibo Emotion Corpus [13], however, relying on word segmentation, not only on acoustic information. Pauses in the voice activity approximate phrase breaks, though the resulting segments may not be as linguistically sound as those derived from word segmentation. However, our segmentation requires no further knowledge and is thus very fast. Though the inclusion of linguistic information has shown to be beneficial [8, 13], we intentionally abstain from using word-based information. Current speech recognition systems are still error-prone [3], especially for arbitrary speech in spontaneous dialogue and complex background conditions as is the case in our showcases. This may have negative influence on the emotion recognition, too. Highly accurate speech recognition for the showcases in CALLAS and IRIS that go beyond keyword spotting is a task for itself that we do not concentrate on. In some showcases, multi-keyword spotting is actually integrated (e. g. E-Tree), but the combination of the linguistic knowledge then takes place at the level of fu-

sion of all input channels (e. g. additionally visual information), not directly with the acoustic information.

3. Strategies to cope with non-prototypical emotions in artistic installations

In previous work (e. g. [1, 2]), a corpus of emotional speech is collected and annotated with emotional states using either emotion categories or emotion dimensions. Typically, the ground truth is given by a majority vote of the labelers and ambiguous cases are usually discarded. In a real-time emotion recognition system, we cannot exclude the occurrence of non-prototypical emotions at runtime. In particular, artistic installations and virtual storytelling environments are characterized by a high degree of uncertainty. Aesthetic experiences may be very different and are hard to predict. For example, for Galassie, we trained a classifier for three emotional states (positive/high-arousal, neutral, negative/low-arousal). In this showcase, the users were intended to control the system via their emotional states as expressed by speech. When analyzing the showcase [5], our project partners identified, however, fourteen different emotion states based on reported user experiences: interest, transport, ludic pleasure, amazement, involvement, creation, serenity, freedom, confusion, irritation, indifference, frustration, boredom, distressed. Interest, transport and ludic pleasure were reported most frequently, that is by 50 % of the users. Of course, it might have been the case that the users were expressing the three trained emotional states when speaking. However, it is very likely that a large number of non-prototypical emotional states occurred during the interaction which were taken as input for our vocal emotion recognition component. This example illustrates that we cannot predict how users interact with the artistic installations and which emotions occur as artistic emotional expression is individual, in particular its strength. We can only define — based on the showcase — which emotions the system will react to, and in order to be able to frequently react to the user's affect we cannot focus on prototypical emotions.

The primary strategy we apply to cope with the diversity of emotions is to train a limited set of emotion classes based on pleasure and arousal in Mehrabian's PAD (**P**leasure, **A**rousal, **D**ominance) model [12]. For instance, five emotion classes (positive/high-arousal, positive/low-arousal, neutral, negative/low-arousal, and negative/high-arousal) were trained for EmoEmma which should then subsume the actually expressed emotions at runtime. In both the E-Tree and EmoEmma scenarios, the classes are mapped onto points in the PAD space. In E-Tree, other modalities also provide PAD values which allows accommodation of non-prototypical emotions by multi-modal fusion of PAD-based emotional representations. Furthermore,

class	recognized as					sum	TP rate
	pos/high	pos/low	neutral	neg/low	neg/high		
pos/high	282	10	46	66	11	415	68.0
pos/low	26	21	18	14	1	80	26.3
neutral	157	11	119	98	14	399	29.8
neg/low	68	6	64	211	13	362	58.3
neg/high	20	4	7	19	38	88	43.2
	553	52	254	408	77	1344	49.9

Table 1. Confusion matrix and true-positive (TP) rates for 5 classes from a single speaker.

a decay is introduced by combining the overall score with previous values to make changes in the tree's appearance smoother and interaction more natural.

Another possibility to cope with non-prototypical data is to concentrate on a few important and specific categories and to add a "garbage" class for all other occurring emotions. This garbage class should then be very general as it has to include very different kinds of emotional expressions. We have not applied this strategy yet in a showcase but we simulate the effects with the following experiment. Table 1 shows confusion matrix and recognition accuracy obtained by 10-fold cross-validation on a database recorded by a single speaker at different points in time with 5 classes: positive/high-arousal, positive/low-arousal, neutral, negative/low-arousal, negative/high-arousal. The difference in frequencies of the classes positive/low-arousal and negative/high-arousal compared to the other classes is due to the data being recorded for two different showcases (E-Tree and EmoEmma) and only one of the showcases made use of all emotions. In the confusion matrix we see that positive/high-arousal is recognized particularly well, while the other classes are most often confused with positive-high. Negative/high-arousal, negative/low-arousal and positive/low-arousal are relatively seldom confused, so in order to have an inhomogeneous garbage class, we train positive/high-arousal and neutral against the rest. Results are shown in Table 2. For comparison, we give results for two well separable classes, negative/high-arousal and positive/low-arousal, against the rest in Table 3. In this case the garbage class is rather homogeneous. Not surprisingly, results are clearly higher for the homogeneous garbage than for the inhomogeneous garbage. In the latter case, the overall result is even lower than for 5 classes. This indicates, that probably there should be more than one garbage class each trained by similar emotional expressions.

As mentioned before, this was a theoretical experiment that has not yet been implemented in a showcase yet, and the definition of the garbage class has been guided by the criterion of homogeneity. However, the recognition of only positive/high-arousal and neutral could be useful in some artistic installations (e. g. Galassie) where users interact just for fun and negative emotions occur rather seldom, and the recognition of only negative/high-arousal and positive/low-

class	recognized as			sum	TP rate
	pos/high	neutral	garb.		
pos/high	301	65	49	415	72.5
neutral	176	167	56	399	41.9
garbage	198	170	162	530	30.6
	675	402	267	1344	46.9

Table 2. Confusion matrix and true-positive (TP) rates with inhomogeneous garbage class: positive/high-arousal, neutral vs. garbage from a single speaker.

class	recognized as			sum	TP rate
	pos/low	neg/high	garb.		
pos/low	42	5	33	80	52.5
neg/high	10	47	31	88	53.4
garbage	136	127	913	1176	77.6
	188	179	977	1344	74.6

Table 3. Confusion matrix and true-positive (TP) rates with homogeneous garbage class: positive/low-arousal, negative/high-arousal vs. garbage from a single speaker.

arousal is conceivable in a scenario, where a reaction should follow to anger caused by the system (e. g. an automated call center), but no extreme positive emotions are expected which however can be interpreted as satisfaction with the system.

A further possibility to deal with a garbage class is multilevel classification by first differentiating between relevant emotions and garbage, and to analyze then which of the relevant emotions actually occurred.

4. Training data for non-prototypical test data

The performance of a classifier on new test data depends strongly on the quality and similarity of the data used to train the classifier. Concerning the showcases we need data with similar speaker groups (some are targeted at selected speakers, others e. g. at children), languages (English by native and non-native speakers, native languages of showcase developers, esp. Finnish and Italian), background noise conditions and occurring emotions. However, currently, there exist no such databases of emotional speech that are suitable to train classifiers for our showcases. In our opinion this is a general problem for speech emotion recog-

nition systems integrated into applications because state-of-the-art technology is not flexible enough to cope with different environmental factors. Existing databases are likewise only suited for their specific conditions. In particular, specific non-prototypical emotions that are usually limited to the application context will rarely be found in existing databases. Thus, application specific training databases have to be recorded.

As mentioned before, EmoVoice was employed in six CALLAS showcases and one IRIS showcase. Due to the large diversity of the showcases, it was not feasible for us to create specific training databases for each showcase. Especially with regard to the number of showcases it was not possible to record as many hours of thoroughly labeled emotional speech data for each showcase as is usually used for offline analysis, not only because the primary goal of the projects was integration into showcases and not extensive data collection but we also assume this again to be a general problem for applied speech emotion recognition under realistic conditions: Though it would be best to record large amounts of data from users interacting with the application, and use them as training data, this is usually not possible because there is no test data yet, it is too time consuming, especially to annotate the data, or not feasible to do by non-experts. For integrated speech emotion recognition systems, application developers need to be able to create databases in a simple and fast fashion on their own. Even if the quality of the databases is not as high as those created by experts, they will be better suited for their purposes. This addresses directly the problem of non-prototypicality in realistic scenarios.

For these reasons, we designed a work flow for the showcase developers to record their own training database adjusted to their showcase. We integrated an easy-to-use interface for recording and training an emotional speech corpus into EmoVoice (see Section 2). The interface offers the possibility to present stimuli that are similar to those occurring in the showcases. The emotion label then results from the stimulus and labeling afterwards is not necessary. The interface lets showcase developers decide on the emotions they want to recognize (though they might not yet know which emotions actually occur, see above), on the language, they can provide as similar as possible background noises and select suitable speakers. One successful method used for emotion elicitation was inspired by the Velten mood induction technique [15] where subjects had to read out loud a set of emotional sentences that should set them into the desired emotional state. However, developers making use of the system were encouraged to change sentences according to their own emotional experiences. E. g. for EmoEmma, the Velten sentences had been completed with actual excerpts from Madame Bovary's dialogues.

The interface allows to quickly build a classifier, but there also arise problems from the recordings being made by non-experts. The Velten method is in principle a very suitable method, but especially when conducted by non-experts it cannot be guaranteed that speakers really immerse in the respective emotions. Thus, even if the recorded emotions may be not fully spontaneous because the sentences are read, they represent a hard and realistic problem because the speakers were no professional actors and did not produce full-blown or prototypical emotions as professional actors would have done. Listening tests on a database recorded for an Italian showcase revealed that it was often hard for humans who could not speak Italian to detect the emotion. Another difficulty arises from the fact that people respond to artwork in a rather individual manner. The analysis of reported user emotions for Galassie provides evidence of the plethora of emotional states people experience when interacting with art. Furthermore, it is hard to control for us whether showcase developers really provided similar settings, and the amount of data is usually small. For these reasons, there is a discrepancy between training and test data, which is likely to occur in real-time systems. Of course, this can seriously affect recognition accuracy. What adds further is that conditions in general are very difficult in some showcases. E. g. background noises in the Interactive Opera are very loud. Especially voices in the background affect the recognition rate badly as the system cannot distinguish which voice it should recognize emotions from.

In order to assess the suitability of the procedure, we analyze data from the EmoEmma, E-Tree and Music Kiosk showcases with 3 classes (positive/high-arousal, neutral, negative/low-arousal) recorded by four male English speakers. Two of them were non-native speakers and class distribution was approximately balanced. The database was recorded with the help of our interface, the stimuli came from the Velten sentences and from sentences occurring in the showcases. Overall recognition accuracies reported in the following were obtained with the Naïve Bayes classifier offline, though speech data and recording conditions were similar to online conditions. Speaker dependent accuracies, that is accuracies that were obtained from each of the four speakers alone by 10-fold cross-validation, ranged from 54.5 % to 65.4 %. When evaluating all speakers together, again in 10-fold cross-validation, accuracy was lower with 49.5 %. Though figures may not sound high for the limited number of classes, they are well above chance level. Furthermore, for good results in online recognition within a showcase the number of classes should generally be limited to two or three. Thus, we can conclude that even if the results obtained from the procedure may not be perfect, it does yield useful results, so that it is a good alternative if no suitable pre-recorded databases exist.

5. Evaluation methods for real-time (artistic) emotion recognition

In general, when evaluating real-time systems, lower recognition rates than for offline analysis have to be expected. The quality of the training database can be assessed with the same methods used for offline analysis, i.e. recognition accuracies in reference to the labels e.g. given by annotators to the emotional events.

However, at run-time there are further problems: A first question is what should be evaluated, the subjective or objective experience. The subjective evaluation can be better or worse than the objective evaluation. In artistic installations, and maybe in other scenarios as well, mainly the subjective experience of the user is of interest and evaluated, i.e. if the user has the impression that the system is responsive. This may diverge from the objective accuracy of the system, though the latter is often difficult to determine, as first a ground truth has to be established. This can be done by annotating test data after run-time though this may be too late and thus not applicable for many purposes. Other methods are physiological measurements as a ground truth (E-Tree) or video observations (ElectroEmotion). However, because of the non-prototypicality of the emotions and as occurring emotions cannot be predicted, it may also be possible to assign an emotional state to a class present in the system only if a garbage class exists. Finally, we need to decide whether to evaluate a system's ability to recognize emotions over time (E-Tree) or whether to evaluate the concept of automated emotion recognition as a whole (EmoEmma). In the first case, we need to compare the system's results with ground truth data at particular points in time. In the second case, we perform a posteriori evaluation of the system usually concentrating on user experience.

In the following, we describe the results of some evaluation studies in more detail. EmoEmma was evaluated by handing each subject a questionnaire about his experience with the interactive character [4]. That is EmoEmma was evaluated as a whole after the user had interacted with it. Questionnaires after the interaction revealed that the subjects responded very positively to the installation and perceived EmoEmma as a believable character that responded appropriately to what they were saying. E.g. on a scale from 0 to 5 they rated 3.6 on average that Emma understood what they were saying. When interpreting this result, we should keep in mind that EmoEmma did not analyze the semantics of the user's utterances, but solely aimed at recognizing the user's emotions from the acoustics of speech. Thus, the result of the user study can be taken as evidence that EmoVoice was effective in the showcase since it was the only mode of interaction.

For E-Tree, we evaluated the performance of the system over a whole interactive session by examining the correlation of arousal detected by multimodal fusion against physiological measurements obtained from Galvanic Skin Response (GSR). We observed a positive linear correlation of 0.79 ($p<0.05$), suggesting that the arousal dimension, at least, of PAD measurement is representative of actual emotional response. However, the contribution of EmoVoice to the detection of arousal was only 11%, so the result can only be taken as evidence that the overall fusion of input channels was appropriate and does not give detailed insight into the individual performance of EmoVoice in the showcase. This is however a further problem in the evaluation of real-time systems: the interplay of many factors such as more than one input modalities or the visual presentation inhibits the isolated inspection of single factors, which is possible in offline analysis.

6. Conclusions

In this paper, we reported on challenges that arise when coping with non-prototypical emotions in the context of artistic installations. When dealing with non-prototypicality, we need to distinguish between the non-prototypicality of stimuli and the non-prototypicality of behaviors. Currently, in CALLAS and IRIS classifiers are trained based on a limited set of standard stimuli validated by psychological research (Velten sentences). At runtime, the users were exposed, however, to a larger and more diverse set of stimuli. Indeed, prototypical stimuli would conflict with the creativity expected from artistic installations. Thus, the next step towards handling non-prototypicality in both research endeavors would be to elicit emotions during training that are as similar as possible to the stimuli at runtime. Currently, CALLAS and IRIS do not rely on manually labeled data for training. Instead the labels are given by user instructions, for example, to read a particular sentence. In order to tackle the non-prototypicality of behaviors, we cannot rely on standard responses to emotional stimuli. Instead the labels should be checked for plausibility and modified accordingly. Furthermore, additional corpora should be collected at runtime as a basis for training new classifiers that are adapted in a better way to the showcase in which they are being used.

Acknowledgments

References

[1] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann. Tales of tuning - prototyping for automatic classification of emotional user states. In Interspeech [7], pages 489–492.

[2] A. Batliner, V. Zeißler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth. We are not amused - but how do you know? User states in a multi-modal dialogue system. In *Proceedings of Eurospeech 2003*, pages 733–736, Geneva, Switzerland, September 2003.

[3] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10–11):763–786, October-November 2007.

[4] M. Cavazza, D. Pizzi, F. Charles, T. Vogt, and E. André. Emotional input for character-based interactive storytelling. In *Conf. on Autonomous Agents and Multiagent Systems*, Budapest, Hungary, May 2009.

[5] G. G. Jacucci, A. Spagnolli, A. Chalambalakis, A. Morrison, L. Liikkanen, S. Roveda, and M. Bertoncini. Bodily explorations in space: Social experience of a multimodal art installation. In *Proc. of the twelfth IFIP conference on Human-Computer Interaction: Interact 2009*. to appear.

[6] S. W. Gilroy, M. Cavazza, R. Chaignon, S.-M. Mäkelä, M. Niranen, E. André, T. Vogt, J. Urbain, M. Billinghurst, H. Seichter, and M. Benayoun. E-tree: emotionally driven augmented reality art. In *Proc. ACM Multimedia*, pages 945–948, Vancouver, BC, Canada, 2008. ACM.

[7] *Proceedings of Interspeech 2005*, Lisbon, Portugal, September 2005.

[8] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transaction on speech and audio processing*, 13(2):293–303, March 2005.

[9] L. A. Liikkanen, G. Jacucci, E. Huvio, T. Laitinen, and E. André. Exploring emotions and multimodality in digitally augmented puppeteering. In *Advanced Visual Interfaces*, Naples, Italy, May 2008.

[10] L. A. Liikkanen and L. Pearce. MusicKiosk: When listeners become composers. An exploration into affective, interactive music. In *Conf. of Music Perception and Cognition*, Sapporo, Japan, August 2008.

[11] M. Lugger and B. Yang. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, USA, March–April 2008.

[12] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, Social, and General Psychology Monographs*, 121:339–361, 1995.

[13] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[14] B. Schuller, R. Müller, M. Lang, and G. Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In Interspeech [7].

[15] E. Velten. A laboratory task for induction of mood states. *Behavior Research & Therapy*, 6(4):473–482, 1968.

[16] T. Vogt and E. André. Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006)*, Genoa, Italy, 2006.

[17] T. Vogt, E. André, and N. Bee. EmoVoice — A framework for online recognition of emotions from voice. In *Proc. Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Kloster Irsee, Germany, June 2008.