An Emotional Storyteller Robot

A. Chella‡, R. E. Barone‡, G. Pilato†, and R. Sorbello‡

‡ DINFO - Dipartimento di Ingegneria Informatica, Università degli Studi di Palermo – ITALY
† ICAR - Istituto di Calcolo e Reti ad Alte Prestazioni Italian National Research Council (CNR)
‡ chella@unipa.it, barone@csai.unipa.it, sorbello_rosario@unipa.it, † g.pilato@icar.cnr.it

Abstract

In this paper we present an "*emotional storyteller robot*": a robot that is capable to tell story in an attractive way using the appropriate affective intonation concerning the piece of story that it's telling.
Our approach is inspired by Taghard's theory about the human ability to find emotional analogies between the situation described in a piece of text and the past experience of the reader in his life.
We use the Latent Semantic Analysis (LSA) for the automatic construction of an "emotional conceptual space" that tries to simulate the emotional-associative capabilities of the human beings.
We also apply the Shroder's theory for the emotional vocalization of text.
At the end of the paper we show the experimental results obtained testing the system on PeopleBot robot.

Introduction

In recent years there has been an increasing interest in the development of *socially interactive robots*, i.e. robots whose main task is the human-robot interaction (e.g a tour guide robot). In this context particular attention has been given to "*socially assistive robots"*, robots that provide assistance through social interaction (Tapus, A., Mataric, M. J. ,and Scassellati, B., March 2007). A multitude of important assistive tasks exist where the social interaction is the main focus: the care of elderly people or the care of individuals with cognitive and social disabilities. In the field of "*socially assistive robots"*, the first effort has been focused on the development of robots designed to be companions to reduce stress and depression (Wada, J., Shibata, T., Saito, T. , and Tanie, K., 2002). Experiments show that feeling of elderly is improved by interactions with the robots (Wada, J., Shibata, T., Saito, T. , and Tanie, K., 2002). A "*socially interactive robot*" must be capable to communicate and interact with people in a "natural manner" according to the human-human communication paradigm. This capability is particularly important in the socially assistive field. We focus our attention on the oral communication and in particular on the emotion expression through voice.

The oral communication is the main channel of communication used by humans. The affective prosody used in a spoken message conveys also information, like the words used by the speaker. As a matter of fact, even in emotionally neutral sentences, speakers provide prosodic information in order to increase the intelligibility of their speech (Berckmoes,C., and Vingerhoets,G., 5, October 2004, Landauer, T. K., Foltz, P. W., and Laham, D., 1998).

In this paper we present an "*emotional storyteller robot*": a robot that is capable to tell story in an attractive way using the appropriate affective intonation concerning the current piece of story that it's telling.

A good storyteller should be capable to choice the appropriate emotional prosody in order to capture the interest of the audience and to make attractive the story reading. Thus to create a storyteller robot two aspects must be considered: the identification of the emotional content of an excerpt and its reading with the most appropriate intonation.

For the identification of emotions from the text we take inspiration from Thagard's theory about Emotional Analogies and Analogical Inference (Thagard, P., and Shelley, C. P., 2001; Barnes, A. and Thagard, P. 1997). The author claims that when a person watches or reads a masterpiece like *Hamlet*, the human capability to feel the same emotions of the main character requires the production of a hidden mapping between the character current situation and some aspects of the actual reader life. According to Thagard's theory our idea is to provide the robot with a sub-symbolic encoding about possible situations and their correlated emotions. To achieve this goal we have built a dataset of texts describing different situations and their correlated emotions, which we have organized in an "emotional semantic space" using the LSA technique.

Latent semantic analysis (LSA) is a well founded technique to represent the meaning of words and to find the conceptual similarity between two pieces of text. According to this paradigm words are roughly coded as vectors in a large dimensional space S (Cullum, J.K., and Willoughby, R.A. 1985; Peters,S. ,Widdows, D., June 2003) in which the semantic similarity between any two chunks of text can be calculated using a properly defined metric. This approach (Quesada, J. F., Kintsch, W., and

Gomez, E., 2003) automatically builds semantic spaces from data in a self-organizing manner and doesn't require any priori theoretical structure construction. It has been empirically established that this methodology is capable to simulate a set of human psycholinguistic cognitive phenomena that involve association capabilities and semantic similarity measures (Agostaro,F., Augello, A. , Pilato,G. , Vassallo, Gaglio,G. S., 2005). In particular Landauer and Dumais (Landauer, T. K., and Dumais, S. T., 1997; Landauer, T. K., Foltz, P. W., and Laham, D., 1998) assert that behind this paradigm there is a fundamental theory for knowledge acquisition and representation. The peculiarities of this technique have been exploited to create a semantic space where the robot can easily find semantic analogies between a piece of text and the previously stored "emotional" knowledge. This empathy capability is used by the robot with the aim of reading the text with the most appropriate prosody according with the inferred emotion.
For the emotional vocal synthesis we refer to the emotional prosody speech study carried out by Schröder (Schröder, M., 2004; Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S , 2001). He has formulated prosody rules that link emotional states with their effects on speech prosody. His model of emotions for speech suggests a representation of emotions in a 3-dimensional space.
The remainder of the paper is organized as follows: first an overview of the system is given; then the process of identification of emotion from the text and the emotional vocal synthesis are described; finally experimental results are shown.

System Overview

We have realized an application for the emotional reading of tales. User interacts with robot through an user friendly interface. He can select a tale among a set of tales stored in the robot. The robot tells the selected story in an emotional way using the most appropriate emotional prosody and displaying emotional avatars corresponding to the emotions evoked by the story.
The developed system is illustrated in Fig 1, which shows also the basic building blocks of the emotional reading system. The "Text segmentation" block subdivides the selected tale in "units of text" that are analyzed by the "Emotion Detection" block in order to extract the emotional content. As a matter of fact, in a tale situations change quickly and often two consecutive sentences concern different scenes. For this reason we have chosen to analyze separately each sentence, which is our basic "unit of text". The "Prosody Assignment" block assigns to each unit of text the appropriate parameters of prosody according to the inferred emotion. The "Emotional Vocal Synthesis " uses these parameters to vocalize the unit of text. Finally the "emotional avatar" displays an emotional avatar corresponding to the inferred emotion.

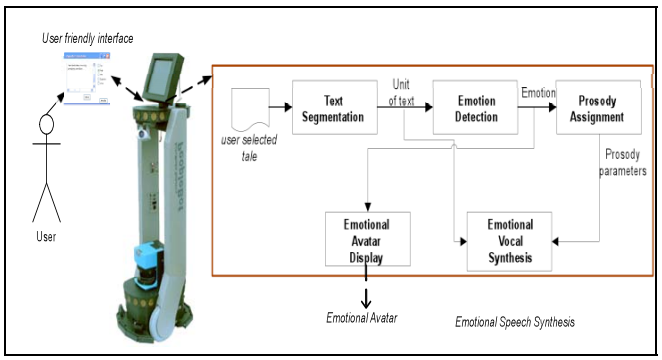

Fig. 1. System Overview

Emotions Detection

The "Emotion Detection" block analyzes each "unit of text" to identify the conveyed emotion.
Different approaches have been used to identify emotions in text (Liu, H., Lieberman, H., Selker, T., 2003; Alm,C. O. , Roth, D., and Sproat, R. 2005; Boucouvalas, A. C. ,2002).
Our approach is inspired by Taghard's theory about empathy and analogy. Taghard (Thagard, P., and Shelley, C. P., 2001) suggests that human empathy aptitude is correlated to the capability to found analogies between situations described in a text and our previous experience. As a consequence the same concept can evoke different emotions according to the person different experiences and culture. Our attempt is to create an "emotional storyteller robot" with its own background of experiences: a robot that exhibits the same human faculty in founding emotional analogies between the piece of text to read and its own "experience".
We have built an "emotional semantic space" during an off-line phase. This space is used on-line by the robot to identify the emotional meaning of the text.
The off-line phase ("emotional space creation") consists of:

- *Semantic Space creation:* We used the Latent Semantic Analysis (LSA) methodology to induce a semantic space S from a collection of data concerning fairy tales.
- *Mapping Emotion in the Space*: We collected a set of texts about typical situations described in tales ("background of experience"). We classified this into "basic" emotion categories according the emotions evoked by the situation described in each text. We coded them in S. The corresponding vectors individuate in S some "areas" corresponding to the "basic" emotion categories.

The on line phase consists of:

- *Emotion Inference:* Each "unit of text" is coded in the space and the geometric closeness between it and the areas corresponding to the "basic"

emotions is evaluated. These quantities indicate the emotional similarity strenght.

In the following subsections we describe the two off-line processes (*Semantic Space creation* and *Mapping Emotion in the Space*) and the on-line process of *Emotion Inference*.

Semantic Space Creation

A set of tales including the Grimm and Anderson's tales has been chosen. The corpus of selected document has been processed in order to remove all words that do not carry semantic information like articles, prepositions and so on (stop words). A corpus of 9375 documents has been obtained and used to build the space. According to the LSA technique a 19847*9375 terms-documents matrix (A) has been created where 19847 is the number of words and 9375 is the number of documents. The TSVD technique, with K = 150, has been applied to A in order to obtain its best approximation. This process leads to the construction of a K dimensional semantic space *S*.

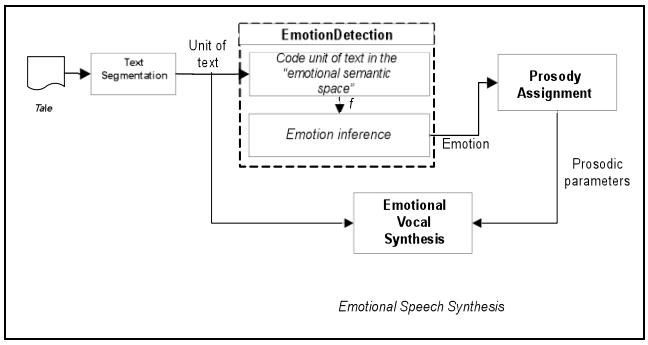

Fig. 2. Emotional Text Reading-Emotion Detection

Mapping Emotion in the Space

English and many other languages contain hundreds of terms that seem to refer emotions. Shaver et al. (Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C, 1987) have analyzed English language and they have sorted affect words using a list of six emotions: *love, joy, surprise, anger, sadness and fear*. We refer to this classification of emotions.

A set of excerpts concerning the "typical" situations described in the tales has been selected and associated with each one E_i of the six emotions proposed by Shaver. For example the text:

"*I went into the forest and I saw the wolf. He wanted devour me all skin, hair, and everything*"

is associated with the "*fear*" emotion.

In this manner we build a background of experience for the robot.

The set of selected excerpts has thus been organized in six subsets of documents each one corresponding to one of the six emotions of Shaver. Each document is coded in S using the folding-in technique (Dumais, S.T., Berry, M.W., and G.W. 0'Brien,1994). According to this procedure the *j-th* document of the subset corresponding to the emotion E_i is coded in *S* as a vector $d_j^{(Ei)}$ obtained summing the vectors representing the terms which compose it. Each emotion E_i is then represented in *S* by the subset of vectors that code the documents associated with it .

Emotion Inference

Each "unit of text" is coded in S as a vector f using the same folding-in technique (Dumais, S.T., Berry, M.W., and G.W. 0'Brien,1994) used for mapping the "emotion prototypes" E_i in S.

The emotional similarity between f and each vector $d_j(E_i)$ is evaluated using the cosine similarity:

$$sim(f, d_j^{(Ei)}) = \frac{f \cdot d_j^{(Ei)}}{\|f\| \cdot \|d_j^{(Ei)}\|}; \qquad (1)$$

An higher value of $sim(f, d_j^{(Ei)})$ is representative of a strong emotional analogy between f and $d_j^{(Ei)}$.

Given an experimentally determined threshold Th, the set M defined as:

$$M = \{ d_j^{(Ei)} \mid sim(f, d_j^{(Ei)}) > Th\} \qquad (2)$$

contains the vectors $d_j^{(Ei)}$ associated with the emotions E_i that have emotional analogies with the input f. If M = ø the input excerpt is labelled as "*neutral*". Otherwise it is labelled with the emotion E_i associated with the vector $d_j^{(Ei)}$ ϵ M having the highest value of $sim(f, d_j^{(Ei)})$.

Prosody Assignment And Vocal Synthesis

We follow the Schröder's theory (Schröder, M., 2004) (Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S , 2001), where a representation of emotions for speech in a 3-dimensional space, given by activation(A), evaluation(E), power(P), has been presented. Schröder proposes some emotional prosody rules that give the possibility to map each point of the space with speech prosody.

Emotion	*A*	*E*	*P*
Neutral	0	0	0
Sad	-8.5	-42.9	-55.3
Angry	34.6	-34.9	-33.7
Fear	31.1	-27.1	-79.4
Joy	28.9	39.8	12.5
Surprise	45,5	15,1	20,3
Love	35	33	7

Table 1-Values of A E and P for the six emotions of Shaver

The six emotions proposed by Shever and the absence of emotion ("*neutral*") have been mapped in this space. For five emotions categories, *neutral, sadness, angry, afraid, happy*, we have used the value of A, E and P defined by Schröder et al. (Schröder, M., Cowie, R., Douglas-Cowie,

E., Westerdijk, M., and Gielen, S , 2001). For the other emotions we have experimentally defined the values of A, E and P according to what illustrated in Table 1.

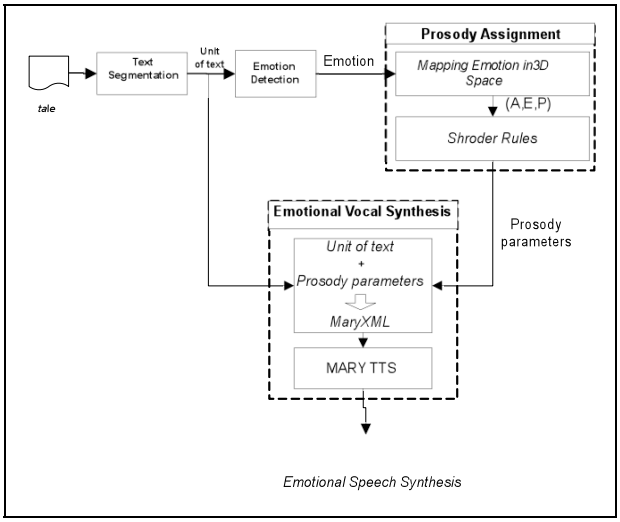

Fig. 3. Emotional Text Reading-Prosody Assignment and Emotional Vocal Synthesis

The module of "Prosody Assignment" receives (Fig. 3), from the "Emotion Detection" module the emotion E_i concerning the analyzed "unit of text" and the value of similarity calculated by(1). A point w (A_w,E_w,P_w) in the 3D-space (A,E,P) is calculated as:

$$\begin{aligned} A_{w=}\ \alpha \cdot A_{Ei};\\ E_{w=}\ \beta \cdot E_{Ei}\ ;\\ P_{w=}\ \gamma \cdot P_{Ei}\ ; \end{aligned} \qquad (3)$$

Where α, β and γ are values proportional to the similarity values calculated for the "unit of text" and (A_{Ei}, E_{Ei} P_{Ei}) is the point in the 3D-space associated with the emotion E_i. The rules defined by Schröder allow mapping this point with the most appropriate speech prosody parameters (pitch, intensity, duration cues in the speech signal and so on). These values and the "unit of text" constitute the input for the "Emotional Vocal Synthesis" module (Fig. 3) that creates a MARYXML (Modular Architecture for Research on speech synthesis XML) document (Fig. 4).

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxmlxmlns="http://mary.dfki.de/2002/MaryXML"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xml:lang="en" version="0.4">
<prosody volume="65" fricative-duration="+18%" plosive-
duration="+18%"liquid-duration="+15%"nasal-
duration="+15%"    vowel-duration="+15%"pause-duration="-
9%"number-of-pauses="+31%"rate="+26%"preferred-boundary-
type="low" accent-prominence="+15%" accent-slope="+37%"
preferred-accent-shape="rising"     range-dynamics="+28%"
range="5.8st" pitch-dynamics="-12%" pitch="+12%">
Hello!
</prosody>
</maryxml>
```

Fig. 4. Example of MARYXML document

This document contains the information for the emotional vocal synthesis. It is sent to the text-to-speech Mary system for the emotional vocalization of the excerpt.
MARY (Schröder, M. and Trouvain, J, 2003) is a platform for research and development on text-to-speech synthesis. It has been originally developed for German and it has been extended to US English. MARY uses an XML language called MaryXML for the representation of its data.

Experimental Results

We have realized an implementation for the PeopleBot robot (Fig.5).

Fig. 5. The PeopleBot robot

The robot shows a list of all its available tales. The user selects a tale using a user-friendly interface and the robot starts the narration.
Fig.6, Table 2 and Table 3 show the results for the incipit of the Grimm's tale "The Frog King". The tale is sent to the "TextSegmentation" module that subdivides it in "unit of text" (fig 4).

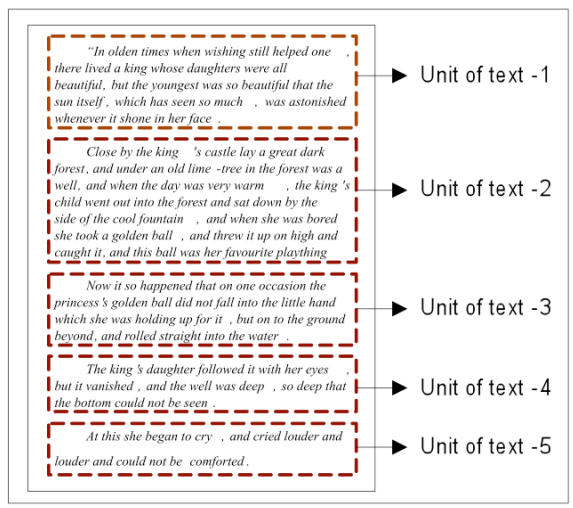

Fig. 6. Text Segmentation

Each unit of text is analyzed by the module of "Emotion Detection" that calculates the concerning emotion and the value of similarity (Table 2). These values are used to calculate the associated point in the 3-dimensional space (A,E,P) (Table 2). We have chosen for (3) α = β= γ=sim(f,djEi).

Unit of text	*Emotion and similarity value*	*(A,E,P)*
Unit of text – 1	E_i= *"Joy"* *Sim = 0,80*	A=0,80*28,9=23,12 E=0,80*39.8=31,84 P=0,80*12,5=10
Unit of text – 2	E_i= *"neutral"*	A= 0 E= 0 P= 0
Unit of text – 3	E_i= *"Surprise"* *Sim = 0,66*	A= 0,66*45,5=30,3 E=0,66*15,1=9,97 P=0,66*20,3=13,40
Unit of text – 4	E_i= *"Surprise"* *Sim = 0,71*	A=0,71*45,5=32,3 E=0,71*15,1=10,72 P=0,71*20,3=14,41
Unit of text – 5	E_i= *"Sad"* *Sim = 0,77*	A=0,77*(-8.5)=-6,54 E=0,77*(-42.9)=-33,03 P=0,77*(-55.3)=-42,58

Table 2 – Emotions, similarity values and A, E and P values for the first five unit of text.

For each unit of text the prosodic parameters are calculated applying the Shroders rules to map the point (A,E,P) with the prosodic parameters (Table3)

Unit of text	*Prosodic Parameters*
Unit of text – 1	pitch = 0.3*A+0.1*E-0.1*P=9,12%; pitch_dynamics = -15+0.3*A-0.3*P=-11,06%; range = 4 + 0.04*A=4,92st; range_dynamics = -40+ 1.2*A+0.4*P=-8,26%; preferred-accent-shape="rising"; accent-slope = 1*A-0.5*E=7,20%; accent-prominance = 0.5*A-0.5*E=-4,36%; preferred-boundary-type = "low"; rate = 0.5*A+0.2*E=17,93%; numb_of_pauses = 0.7*A=16,18%; pause_duration = -0.2*A=-4,62%; vowel_duration = 0.3*E+0.3*P=12,55%; nasal_duration = 0.3*E+0.3*P=12,55%; liquid_duration = 0.3*E+0.3*P=12,55%; plosive_duration = 0.5*A-0.3*E=2,01%; fricative_duration = 0.5*A-0.3*E=2,01%; volume = 50+0.33*A=57,63%;

Table 3. Prosodic Parameters for the unit of text 1

The prosodic parameters are used to obtain a MARYXML document (Fig. 7) that is sent to the MARY TTS for the emotional vocalization.

Conclusions

We have presented an emotional storyteller robot with emotional associative capabilities provided by a geometric and sub-symbolic model. We are thinking to conduct experiments involving also Psychologists. We believe that this study can be extended to obtain a socially assistive robot. Future work will concern the development of a speech recognition ability for the robot that allows it to receive vocal commands. We are also involved in the development of a system for the emotion recognition that allows the robot to receive emotional feedback from its audience.

```
<prosody pitch="9.120000000000001%" pitch-dynamics="-
11.064%" range="4.9248st" range-dynamics="-
8.256%"preferred-accent- shape="rising" accent-
slope="7.200000000000001%" accent-prominence="-
4.359999999999999%" preferred-boundary-type="low"
rate="17.928%" number-of-pauses="16.184%" pause-
duration="-4.6240000000000006%" vowel-duration="12.552%"
nasal-duration="12.552%" liquid-duration="12.552%"
plosive-duration="2.008000000000001%" fricative-
duration="2.008000000000001%" volume="57.6296">
In olden times when wishing still helped one, there
lived a king whose daughters were all beautiful, but the
youngest was so beautiful that the sun itself, which has
seen so much, was astonished whenever it shone in her
face.
</prosody></maryxml>
```

Fig. 7. MARYXML document

References

Agostaro,F., Augello, A. , Pilato,G. , Vassallo, Gaglio,G. S., 2005. A Conversational Agent Based on a Conceptual Interpretation of a Data Driven Semantic Space, Lecture Notes in Artificial Intelligence, Springer-Verlag GmbH, vol. 3673/2005, pp 381-392

Alm,C. O. , Roth, D., and Sproat, R. 2005, Emotions from text:machine learning for text-based emotion prediction, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing,* Vancouver, British Columbia, Canada pp: 579 – 586

Barnes, A. and Thagard, P. 1997, Empathy and analogy. *Dialogue: Canadian Philosophical Review,* 36: 705-720

Berckmoes,C., and Vingerhoets,G., 5, October 2004, Neural Foundations of Emotional Speech *Processing Current Directions in Psychological Science,* Volume 13—Number , pp. 182-185

Boucouvalas, A. C. ,2002. Real Time Text-to-Emotion Engine for Expressive Internet Communications, Chapter in Volume 5 *Emerging Communication: Studies on New Technologies and Practices in Communication*. Edited by: G. Riva , F. Davide and W. IJsselsteijn, pp. 305-318.

Cullum, J.K., and Willoughby, R.A. 1985. Real rectangular matrices, in Lanczos Algorithms for Large Symmetric Eigenvalue Computations-Vol. 1, Theory. Boston, MA: Brickhauser

Dumais, S.T., Berry, M.W., and G.W. 0'Brien,1994. *Using linear algebra for intelligent information retrieval*, 573-595

Feil-Seifer and Mataric',M. J., June 28-July 1, 2005, Defining Socially Assistive Robotics In Proceedings of the 9th International Conference on Rehabilitation Robotics, Chicago, IL, USA,

Landauer, T. K., and Dumais, S. T., 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of

the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T. K., Foltz, P. W., and Laham, D., 1998, Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

Liu, H., Lieberman, H., Selker, T., 2003. A Model of Textual Affect Sensing using Real-World Knowledge. *Proceedings of the 2003 International Conference on Intelligent User Interfaces*, IUI 2003, January 12-15, 2003, Miami, FL, USA. ACM 2003, ISBN 1-58113-586-6, pp. 125-132. Miami, Florida

Peters,S. ,Widdows, D., June 2003. Word vectors and quantum logic experiments with negation and disjunction In *Mathematics of Language*, 8, Bloomington, Indiana

Quesada, J. F., Kintsch, W., and Gomez, E., 2003. Latent Problem Solving Analysis as an explanation of expertise effects in a complex, dynamic task. In R. Alterman & D. Kirsh (Eds.), Proceedings of the 25th Annual Conference of the Cognitive Science Society. Boston, MA. 0-8058-4581-X

Jan van Santen et al. 2003. Applications of computer generated expressive speech for communication disorders. In *EUROSPEECH 2003*, 1657–1660.

Schröder, M. and Trouvain, J, 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching, *International Journal of Speech Technology*, 6, pp. 365-377.

Schröder, M., 2004. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions, *Proc. Workshop on Affective Dialogue Systems: Lecture Notes in Computer Science*, Kloster Irsee, Germany, 2004, pp. 209–220..

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S , 2001. Acoustic correlates of emotion dimensions in view of speech synthesis, *Proceedings of Eurospeech 2001* Aalborg, Denmark, pp. 87–90

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C, 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52, 1061–1086

Tapus, A., Mataric, M. J. ,and Scassellati, B., March 2007. Socially Assistive Robotics, *IEEE Robotics & Automation Magazine*

Thagard, P., and Shelley, C. P., 2001. Emotional analogies and analogical inference. In *D. Gentner & K. H. Holyoak & B. K. Kokinov (Eds.), The analogical mind: Perspectives from cognitive science* (pp. 335-362). Cambridge, MA: MIT Press.

Wada, J., Shibata, T., Saito, T. , and Tanie, K., 2002. Analysis of factors that bring mental effects to elderly people in robot assisted activity. In Proc 2002 IEEE/RSJ Int. Conf. Intelligent Robots and Systems EPFL, Lausanne, Switzerland, , pp. 1152–1157.