A Methodology for Reader's Emotional State Extraction to Augment Expressions in Speech Synthesis

Dimitrios Tsonos
Department of Informatics and Telecommunications, University of Athens, Greece
ea02534@di.uoa.gr

Gerasimos Xydas
Department of Informatics and Telecommunications, University of Athens, Greece
gxydas@di.uoa.gr

Georgios Kouroupetroglou
Department of Informatics and Telecommunications, University of Athens, Greece
koupe@di.uoa.gr

Abstract

This paper presents a methodology for the real-time extraction of readers' emotional state from documents as well as the representation of emotionally annotated documents into acoustic modality. Using the Self Assessment Manikin Test we performed preliminary experiments on the extraction of Pleasure – Arousal – Dominance (P.A.D.) annotation rules that the documents evoke to the readers. The rules are used in an automated procedure that assigns text's formatting and structure of documents (meta-data) to emotional state values. During the vocalization of documents, these values, and consequently the documents' meta-data, are carried by different expressions of speech using text-to-speech synthesis. The proposed system architecture is language independent and content-free.

Keywords*: emotional state extraction–annotation of documents, expressive speech synthesis, document-to-audio, speech representation of text formatting,*

1. Introduction

Following the wide penetration of the internet and the computer technology in the society, electronic documents are nowdays recognized as a widespread medium for transmitting information like ideas, news, messages, discussions, etc. In electronic documents, e.g. web pages, e-books and on-line magazines, the authors utilize structural and layout elements (meta-data) to personalize the content. The style of a document sets the typographic rules, used more often in technical or scientific papers and reports than in magazines. There are several studies that reveal the coherence of meta-data with the readability of a document [1] [2] and the meanings that the reader comprehends [3] [4].

Emotions and the emotional state of the reader depend on document's structure, layout and text formatting. Multiple combinations of colors [6], font size, type and style in a document affects the emotional state [7] [8] and consequently the readability of the document [9] [10] not only in printed but also in electronic format [5].

Studies on the Human Computer Interaction field focus on the user's emotional response during the interaction [11] [12] [13]. A common used experimental procedure is the Self Assessment Manikin (S.A.M.) test, introduced in 1985 by P. J. Lang [36]. Using pictorial questionnaires, participants of the experiment have to assess their emotional state on three dimensions, namely Pleasure, Arousal and Dominance (also known as P.A.D. approach) [14]. S.A.M. procedure came up as a test for the assessment of advertisements [15]. It is a cross – cultural and language independent test [16]. Laarni et al. performed this kind of tests for the evaluation of readability of documents [9] and web pages [17].

In a similar way, acoustic stimuli affect the emotions and the emotional state of the listener. Expressive Speech Synthesis (E.S.S.) [18] is a method for conveying emotions through speech, using the variations and differences of speech characteristics. Studies focus on specific emotions that can be extracted from speech characteristics [19] [20] [21]. Furthermore, in E.S.S., the term "variety of styles" is used as a domain dependent point of study. The styles can be "good-bad news", "yes-no questions" [22] "storytelling" [23] "military" [24] etc. Emotional prosody is language and culture depended [25]. Thus the research of emotional prosody variations has to be conducted in multiple languages and modeled for each language separately.

There have been studies on modeling expressive speech using the Pleasure-Arousal-Dominance approach [26] [27]. The advantage of using this method is that the values of P.A.D. dimensions are continuous. Through an emotional state it is possible to map P.A.D. values in a

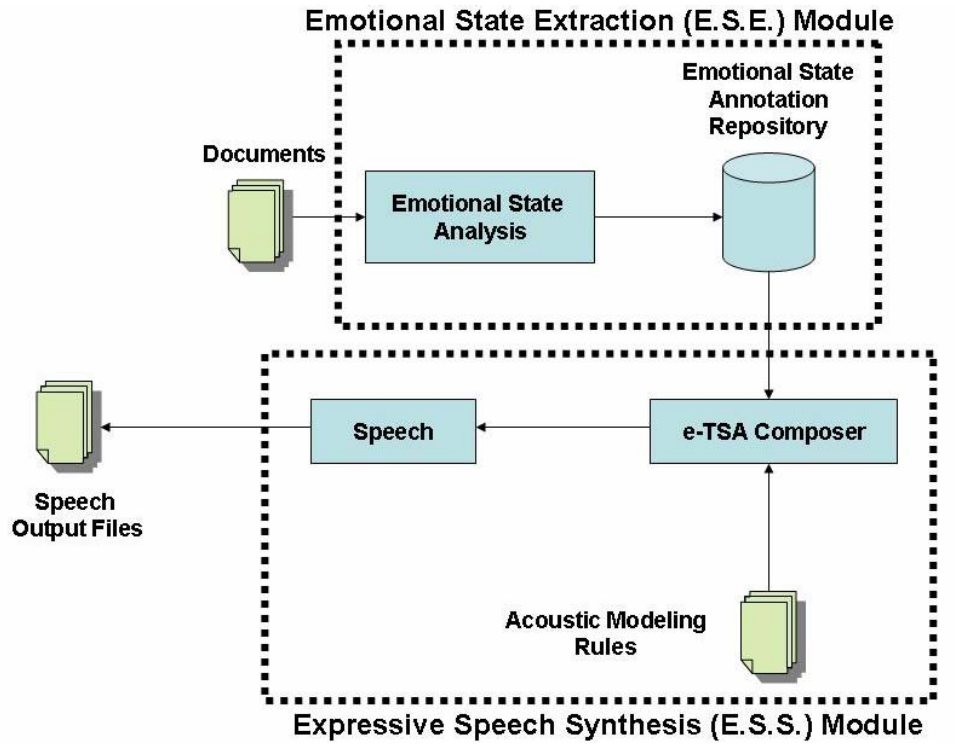

Figure 1. The proposed XML-based Architecture.

specific emotion (or variations of the emotion). For example, the emotion "happy" can have variations like "quite happy", "very happy", "less happy".

Emotional Text Analyzer [28] illustrates an approach for opinion mining - textual affect sensing based on the modified version of lexical affinity. In [35] the Emotion Estimation module is used for the automatic extraction of affective content of the textual messages in a chat system.

Conveying information (meta-data) that accompanies a text from visual modality to acoustic is a complex procedure. Previous works have accomplished this "conversion" using methods like semantics extraction from documents and convey the hidden logic into acoustic modality [29] [30] [31]. However, there is not a systematic work on the creation of an automated reader's emotion – emotional state extractor that bears from text formatting and the structure of documents. Most approaches are concerned with extracting specific emotions from lexical elements.

In this work we propose a methodology for the real-time extraction of readers' Emotional State (E.St.) from documents' metadata and the P.A.D. annotation. Using the E.St. annotated document, we are able to map the states into variations and differences of speech characteristics. The proposed architecture is language-free and document domain independent.

Also, a preliminary test of our approach is performed using specific structural elements of a document.

2. System's Overview

In this study we propose an XML-based system that automatically extracts the reader's emotional state transitions from documents, in real-time mode and conveys them into the acoustic modality by the means

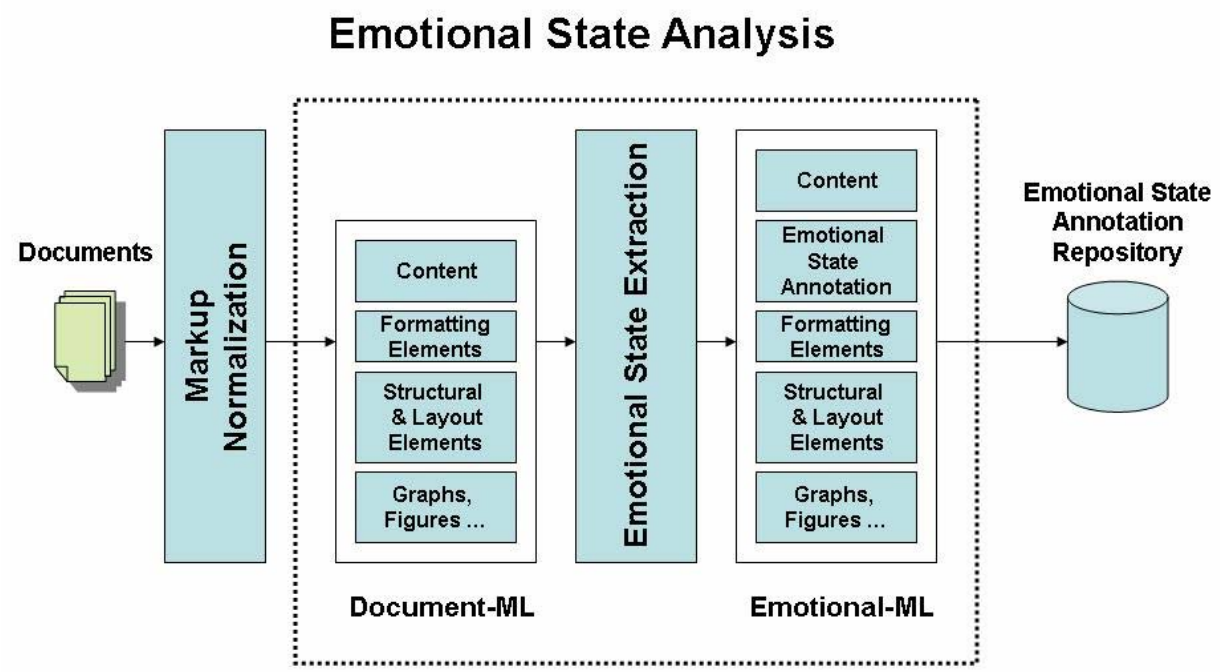

Figure 2. The Emotional State Extraction Module.

of expressions during speech synthesis. The proposed architecture:

a) produces annotated documents with supplementary information about the Pleasure, Arousal and Dominance that bear from documents' structure and text formatting.

b) maps the P.A.D. values into speech prosodic variables in order to convey the hidden emotional state information of documents into speech, using the methodology proposed in [27].

In Figure 1 the system's architecture is presented. The system is divided into two modules:

- Emotional State Extraction (E.S.E.) module.
- Expressive Speech Synthesis (E.S.S.) module.

In E.S.E. module, the Pleasure, Arousal and Dominance annotation is obtained, preserving the original documents' annotation. The resulting XML documents are stored in a database for quick access by the E.S.S. module or by an external third-party application.

Having the documents "emotionally" annotated and stored into the annotation repository, the E.S.S. module (using the e-TSA platform [30]) is able to use the content together with accompanying emotional state meta-data.

3. Emotions Extraction from Documents

3.1. Initial Annotation of Documents

The E.S.E. Module (Figure 2) is able to process any XML-based document. The Markup Normalization Module converts whatever tagged document into the DAISY/NISO compliant format proposed in [29]. This format specifies the following meta-data:

- Text Formatting meta-data. (i.e. bold, italics, font size),
- Text Structure meta-data. (i.e. chapter, title, paragraph),
- Text Layout meta-data. (i.e. like columns, headlines, borders),
- Non-textual meta-data (i.e. Figures, Drawing, Pictures, Logos).

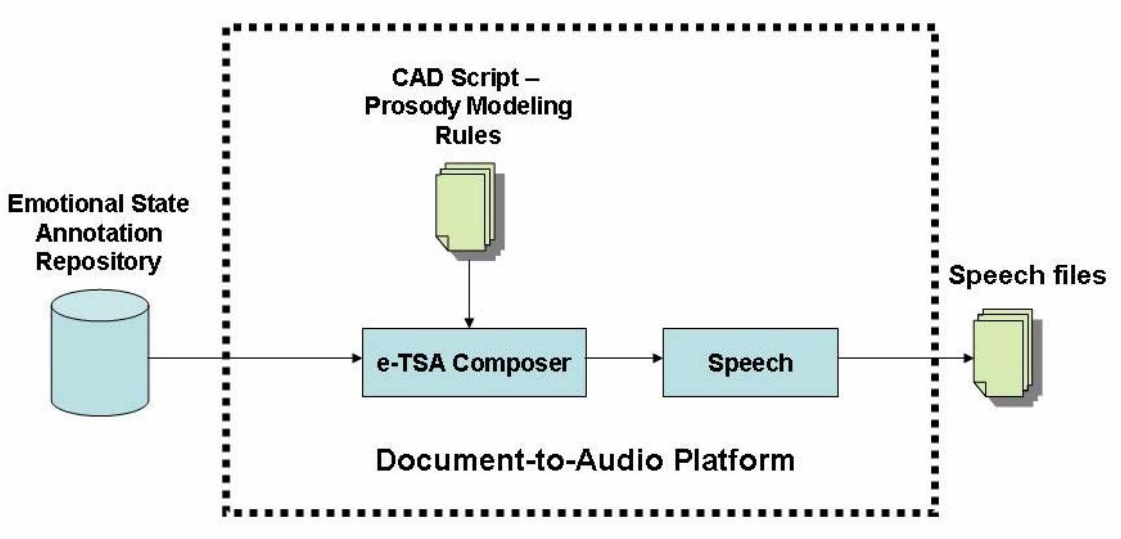

Figure 3. The Expressive Speech Synthesis Module.

These meta-data are host in the tagged Document-ML, which is processed in a later stage, by the E.S.E. Module for the Emotional State Annotation.

3.2. Emotional State Annotation

Findings in [9] showed that part of the above mentioned visual elements affect directly the emotions, the emotional state of readers and consequently the readability of the document. Here, we propose a more systematic study of documents' meta-data in this direction. We investigate the most common used elements and combinations of them. These elements are:

- Text formatting meta-data (bold, italics, underlined, font type and size, color).
- Text structure meta-data. In this case the most common elements will be used, such as different level of title-subtitle, caption, footnote-note, author, references, etc. There is a limitation in the study of the reader's emotional response and the readability of document including these elements because they are documents' domain dependant.

Adding text layout meta-data increases the complexity of our study, so they will be excluded.

The completion of the experimental procedure result the mapping rules for the P.A.D. annotation of documents. The rules are used in the E.S.E. module which produces an XML file (Emotional-ML) that is stored in a database called Emotional State Annotation Repository.

4. Expressive Speech Synthesis and Meta-Data

In system's architecture, the new format of the document is the Emotional-ML, stored in a database called Emotional State Annotation Repository. In the literature and in the preliminary tests, the emotions and the emotional state that are extracted from documents' meta-data, are *continuum nonextreme*. While reading document the emotional state changes over time according to the presented visual elements of the document (continuum). In most cases there are not extreme variations in the use of visual elements (nonextreme). For example, as presented in the preliminary tests, the "Title" is in bold, 14pt Times New Roman and the "Author" is in bold, 12pt. Times New Roman. Using P.A.D. emotional dimensions, and not specific emotions, we are able to fulfill these requirements.

Unit selection from the existing emotional speech database recordings, limits the number of emotional states to be represented because most databases are structured using specific emotions [34]. In our methodology, we need a continuum speech synthesis approach. This is achieved using emotional speech prosody rules as proposed in [27].

The study in [27] is conducted in German and English language. There was an assumption using [32]

that vocal emotion expression is very similar across languages in contraposition in [25] that some language effects are reported. In our study we aim to evaluate these rules in Greek language so to make the appropriate corrections (if needed) in order to adapt the model in Greek language and if it is necessary to use the same methodology for the creation of emotional prosody rules in the Greek language.

The advantage of this model, and generally the described methodology, is the immediate correlation of Pleasure – Arousal – Dominance (named also as Evaluation – Activation – Power respectively) with the prosody variations. This correlation gives us the opportunity to map the documents' metadata into speech elements using the P.A.D. dimensions as medium.

In Figure 3 the E.S.S. Module is described. The mapping is performed using features from the Document-to-Audio platform (DtA) [30]. The mapping is done according the information provided by the Cluster Auditory Definition (C.A.D.) scripts. Using C.A.D. scripts for the implementation of the rules, we are able to map the P.A.D. values to prosodic variations.

5. Self Assessment experimental procedure

As above mentioned, the Self Assessment Manikin (S.A.M.) Test is used in a standardized experimental procedure using the guidelines presented in [14]. The participants are asked to assess their emotional state using the provided P.A.D. (9-scale) pictorial questionnaire. The participants are adults both sexes, aging from 18 – 40 year old, with different educational background.

In the first session of the procedure, the P.A.D. assessment is done in variations of discrete meta-data (e.g. how the different sizes, of specific font type affect the P.A.D.). In the second session, the most common combinations of text's formatting and structure in different type of documents is given. This reveals the dependencies of the variations of each meta-data and their combination.

It is worth noticing that the content of the document should not affect the emotional state of the participants. Hence, emotionally neutral content is provided and before the experimental procedure has to be noticed that each participant should not be affected by the content but only by the visual information of the text. The results reveal the link between the meta-data's variations and the P.A.D. percentage variations.

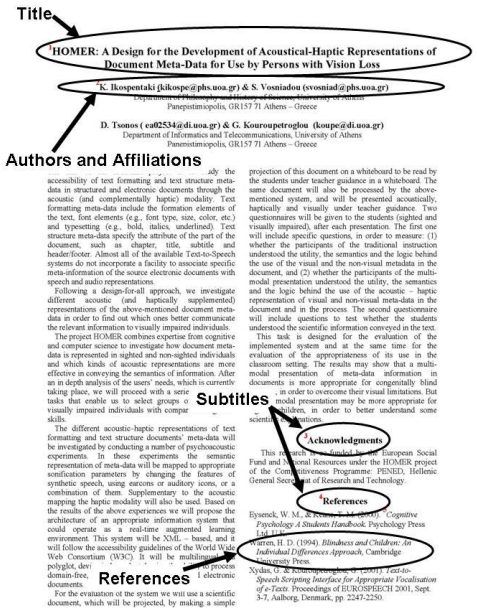

Figure 4. The visual stimulus of the experimental procedure

6. Preliminary Tests for Structural meta-data

Due to the originality of our study in the Human Computer Interaction and Information Extraction approaches, we conducted a preliminary test for the propriety of our experimental procedure.

The visual stimulus of the experimental procedure was a one-page, two-column short scientific paper (Figure 4) containing the following elements:

- Title of the abstract, using text in bold, 14pt. Times New Roman.
- Authors and Affiliations in bold, 12pt. and 11pt. respectively, Times New Roman.
- Subtitles in bold, 12pt. Times New Roman.
- Body of text, in normal, 10pt., Times New Roman.
- References in 10pt. Times New Roman with name of study in italics.

Table 1. The Self Assessment results of the preliminary test. The three dimensions are numbered from -100% to +100% variations. The Pleasure – Arousal values is possible to be mapped in specific emotions.

Structural Elements	Pleasure (%)	Arousal (%)	Dominance (%)	Emotion
Title	+12	+28	+56	Interested - Curious
Author	0	+8	+12	Excited - Reverence
Subtitle 1	+4	-16	-12	Repentant - Apathetic
Subtitle 2	+12	-12	+12	Repentant - Apathetic
Reference	+28	-48	+24	Confused - Overwhelmed

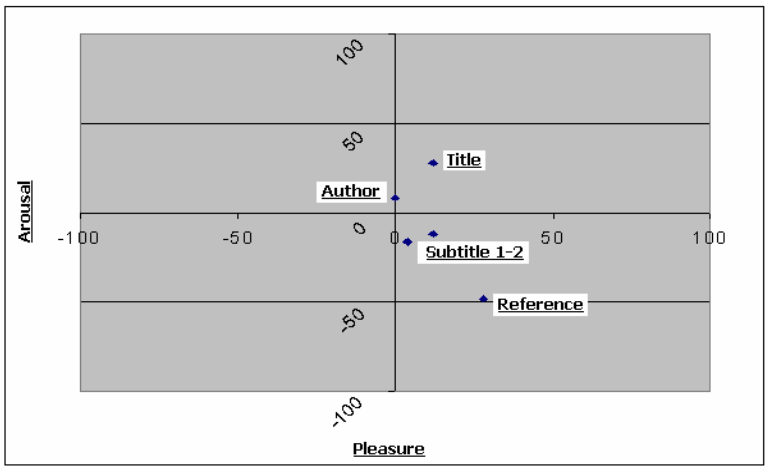

Figure 3. The Pleasure – Arousal variations using percentage diagram for their emotional mapping.

Six male and six female, from 24 to 33 year-old, having high or higher educational background, speaking English fluently participated in the experiment. The participants were asked to assess the Title, Subtitles, Authors and the title of the references. The body of text was considered as baseline to our study (emotionally neutral). The results are presented in Table 1.

Using the Pleasure – Arousal axis the results are shown in Figure 3. Applying these two axis, in percentage scale from -100% to +100%, we are able to appear the verbal expression of the emotions as presented in [33] by K. Scherer. The starting - zero point of the axis is considered as the base line (emotionally neutral).The emotional state of the participants can be mapped in specific emotions. The extracted verbal expressions of the emotions are presented in Table 1.

The results were encouraging and quite expectable from our day-life experience. It is also notable that the preliminary tests have confirmed that the use of italics is less readable [9].

7. Conclusions – Future Work

We have presented an approach for the automatic emotional state annotation of documents using the P.A.D. dimensions. The proposed methodology maps documents' meta-data into speech prosodic variations.

The output of the Emotional State Extraction Module can be utilized in different Expressive Speech Synthesis systems. In our study, we selected the E.S.S. using the P.A.D. mapping, but also E.S.S. systems using emotional or speech style prosodic rules can be used, due to the corresponding of P.A.D. values into specific emotions.

Moreover, the results and the models will indicate which the most appropriate E.S.S. technique is. We should take into account that many variations in speech may confuse the listener during the content's comprehension. Limiting the emotional state variations and consequently the prosodic variations, unit selection

technique would be more appropriate, increasing the naturalness of the speech output.

We intend to extend the experimental procedure in a dataset not containing only the structural elements as in the preliminary tests, but also using text formatting elements. The evaluation procedure of our methodology – approach aims:

- To reveal the readability of the documents
- On how the reader comprehends and better understand the content of the document using different combinations of text formatting and text structure elements.

Also we plan to extend our work by:

1. studying the variation of prosodic elements in expressive speech synthesis for non-sighted participants and how they comprehend the given content and
2. studying how the combination of document's meta-data and emotionally non-neutral content of document, can be conveyed from visual to acoustic modality, in order to have the same affect to the listener as to the reader.

8. Acknowledgements

The work described in this paper has been co-funded by the European Social Fund and Hellenic National Resources under the HOMER project of the Competitiveness Programme: PENED, Greek General Secretariat of Research and Technology.

9. References

[1] K. Holmqvist, J. Holsanova, M. Barthelson and D. Lundqvist, "Reading or scanning? A study of newspaper and net paper reading", In Hyönä, J. R., and Deubel, H. (Eds.)*, The mind's eye: cognitive and applied aspects of eye movement research,* Elsevier Science Ltd, 2003, pp. 657-670.

[2] K. Holmqvist and C. Wartenberg, "The role of local design factors for newspaper reading behaviour – an eye-tracking perspective. Daily newspaper layout-designer's prediction of reader's visual behaviour - a case study", *Lund University Cognitive Studies*, 127, Lund, LUCS 2005.

[3] N. Küpper, "Recording of Visual Reading Activity: Research into Newspaper Reading Behaviour." *(Available as pdf from http://calendardesign.de/leseforschung eyetrackstudy.pdf),* 1989.

[4] Holmberg N., *Eye movement patterns and newspaper design factors. An experimental approach*, Master Thesis, Lund University Cognitive Science, Sweden, 2004.

[5] K. Larson, "The Technology of Text", *IEEE Spectrum*, May 2007. pp 20 – 25.

[6] Barrien F., *Color and Human response*, Van Norstrand Reinhold, New York , 1978.

[7] J.A. Sánchez, N. P. Hernández, J. C. Penagos and Y. Ostróvskaya, "Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states". *Proceedings of the Symposium on Human Factors in Computer Systems IHC 2006*, Natal,Brazil. 2006.

[8] J.A. Sanchez, I. Kirschning, J.C. Palacio, Y. Ostróvskaya, "Towards mood-oriented interfaces for synchronous interaction", *Congreso Latinoamericano de Interacción Humano-Computadora (CLIHC'05),* Cuernavaca, México, 2005, pp.1-7.

[9] J. Laarni, "Effects of color, font type and font style on user preferences", *In C. Stephanidis (Ed.) Adjunct Proceedings of HCI International 2003*, Heraklion, Crete, Lawrence Erlbaum Associates Inc, 2003, pp. 31-32.

[10] T. Saari, M. Turpeinen, J. Laarni, N. Ravaja and K. Kallinen, "Emotionally Loaded Mobile Multimedia Messaging", *Entertainment Computing - ICEC 2004, Third International Conference*, Eindhoven, The Netherlands, September 1-3, 2004, pp. 476-486.

[11] The Humaine Portal, *http://emotion-research.net/*

[12] C. Dormann, "Affective experiences in the home: measuring emotions", *HOIT2003*, California, U.S.A., April 6-8, 2003.

[13] L. Kärkkäinen, and J. Laarni, "Designing for small display screens", *NordiCHI 2002*, Aarhus, Denmark, October 19-23, 2002, pp. 227-230

[14] Lang P.J., M. Bradley and B. Culthbert, *International Affective Picture System (IAPS): Instruction Manual and Affective Ratings*, Technical Report A-6, The Center for Research in Psychophysiology, University of Florida, U.S.A., 2005.

[15] R. P. Bagozzi, M. Gopinath and P. U. Nyer, "The Role of Emotions in Marketing", *Journal of the Academy of Marketing Science,* Vol. 27, 1999, pp. 184 - 206.

[16] J. D. Morris, "Observations SAM: The self-assessment manikin- An efficient cross-cultural measurement of emotional response", *Journal of Advertising Research,* November-December, 1995, pp. 63-68.

[17] L. Kärkkäinen and J. Laarni, "A New Test for Web Page Evaluation", *Proceedings of the IADIS International Conference WWW/Internet 2003,* ICWI2003, Algarve, Portugal, November 5-8, 2003, pp. 1255-1256.

[18] "Special Section on Expressive Speech Synthesis", as presented in the Editorial of *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no 4, July 2006, pp. 1097-1098.

[19] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny and J. Pitrelli, "A Corpus-Based Approach to <ahem/> Expressive Speech Synthesis", *5th ISCA Speech Synthesis Workshop* 14th-16th June 2004, Carnegie Mellon University Pittsburgh, pp. 79-84

[20] C. Drioli, G. Tisato, P. Cosi and F. Tesser, "Emotions and voice quality: experiments with sinusoidal modelling", *In VOQUAL'03*, Geneve, Switzerland, August 27-29, 2003, pp. 127-132.

[21] C.-H. Wu, C.-C. Hsia, T.-H. Liu and J.-F. Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, No 4, 2006, pp. 1109-1116.

[22] Pitrelli J.F., Bakis R., Eide E. M., Fernandez R., Hamza W.; Picheny M.A., "The IBM expressive Text-to-Speech synthesis system for American English", *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 14, Iss.. 4, July 2006, pp. 1099- 1108.

[23] Theune M.; Meijs K.; Heylen D. ;Ordelman R, "Generating expressive speech for storytelling applications", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, Iss. 4, July 2006, pp. 1137-1144.

[24] W.L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters", *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, Santa Monica, USA, September 11-13, 2002, pp.163-166

[25] F. Burkhardt, N. Audibert, L. Malatesta, O. Türk, L. Arslan and V. Auberge, "Emotional Prosody - Does Culture Make A Difference?", *Speech Prosody 2006,* Dresden, Germany, May 2-5, 2006.

[26] A. Iida, N. Campbell, F. Higuchi, F. and M. Yasumura, "A corpus-based speech synthesis system with emotion", *Speech Communication*, Elsevier, vol. 40, no 1, April 2003, pp. 161-187.

[27] Schroeder M., "Expressing degree of activation in synthetic speech" *IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, Iss. 4, July 2006, pp. 1128-1136*

[28] Emotional Text Analyzer, *http://emotion.informatik. uni-augsburg.de/*

[29] D. Tsonos, G. Xydas, and G. Kouroupetroglou, "Auditory Accessibility of Metadata in Books: A Design for All Approach", *Lecture Notes in Computer Science (LNCS)*, Springer, Vol. 4556, pp. 436-445, 2007.

[30] G. Xydas, V. Argyropoulos, T. Karakosta and G. Kouroupetroglou, "An Open Platform for Conducting Psycho-Acoustic Experiments in the Auditory Representation of Web Documents" , *in Proceedings of the ACOUSTICS 2004 Conference*, Thessaloniki, Greece, September 27-28, 2004, pp. 157-164.

[31] D. Spiliotopoulos, G. Xydas, G. Kouroupetroglou and V. Argyropoulos, "Experimentation on Spoken Format of Tables in Auditory User Interfaces", Universal Access in HCI, *Vol. 8, Proceedings of the 11th International Conference on Human-Computer Interaction (HCII-2005)*, Las Vegas, USA, July 2005, pp. 22-27.

[32] K. R. Schrerer, R. Banse, and H.G. Wallbott, "Emotion Inferences from vocal expression correlate across languages and cultures", *J. Cross-Cultural Psychol.*, vol.32, no. 1, 2001, pp.76-92.

[33] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Science Information*, 44(4), 2005, pp. 693-727.

[34] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods", *Speech Communication*, Vol. 48, no 9, January 2006, pp. 1162-1181.

[35] C. Ma, H. Prendinger and M. Ishizuka, "A Chat System Based on Emotion Estimation from Text and Embodied Conversational Messengers", *Proceedings of 4th International Conference Entertainment Computing - ICEC 2005*, Sanda, Japan, September 19-21, 2005, p.p. 535-538.

[36] P. J. Lang, "The cognitive psychophysiology of emotion: Fear and anxiety" in A. Tuma & J. Maser (Eds.), *Anxiety and the anxiety disorder*, Lawrence Erlbaum, 1985, pp. 131-170.