Emotional Prosodic Model Evaluation for Greek Expressive Text-to-Speech Synthesis

Dimitrios Tsonos[1], Pepi Stavropoulou[1], Georgios Kouroupetroglou[1], Despina Deligiorgi[2], and Nikolaos Papatheodorou[1]

[1] National and Kapodistrian University of Athens, Department of Informatics and Telecommunications, Athens, Greece
{dtsonos,pepis,koupe}@di.uoa.gr
[2] National and Kapodistrian University of Athens, Department of Physics, Athens, Greece
despo@phys.uoa.gr

Abstract. In this study we introduce a novel experimental approach towards the evaluation of emotional prosodic models in Expressive Speech Synthesis. It is based on the dimensional emotion expressivity and adopts the Self-Assessment Manikin Test. We applied this experimental approach to evaluate an emotional prosodic model for Greek expressive Text-to-Speech synthesis. We used two pseudo-sentences for each of the Greek and English HMM-based synthetic voices, implemented in the MARY TtS platform. Fifteen native Greek participants were asked to assess eleven emotional states for each sentence. The results show that the "Arousal" dimension is perceived as intended, followed by the "Pleasure" and "Dominance" dimensions' ratings. These preliminary findings are consistent with the results in previous studies.

Keywords: Expressive Speech Synthesis, prosody evaluation, Text-to-Speech, emotional state.

1 Introduction

Expressive Speech is "the speech which gives us information, other than the plain message, about the speaker and triggers a response to the listener" [1]. Accordingly, Expressive Speech Synthesis (ESS) [2] is a method for conveying emotions (and other paralinguistic information) through speech, using the variations and differences of speech characteristics. There is a plethora of studies towards the creation of expressive speech synthesis in order to achieve a more natural result during Human-Computer Interaction. Furthermore, in the domain of document accessibility, emotional-based mapping has been recently proposed [3] for rendering document signals to the auditory modality through Document to Audio systems [4, 5] that incorporate ESS. This approach aims to overcome the limitation of the current Text-to-Speech (TtS) systems [6] towards an effective acoustic provision of the semantics and the cognitive aspects of the visual (such as the typographic signals) and non-visual (such as the logical structure) knowledge embedded in rich text documents [7].

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2014, Part II, LNCS 8514, pp. 166–174, 2014.

Several works [8-11] suggest that there is a certain universal character behind the vocal expression of emotions. The approach followed by Schröder [11] for speech rule-base synthesis builds on the hypothesis that vocal emotion expression is very similar across languages. More specifically, he presents a model for expressive speech synthesis in MARY TtS [12] using the dimensional "Pleasure", "Arousal" and "Dominance" (PAD) approach of emotions. PAD methodology has the advantage of using continuous values of the emotional expression. PAD values can be mapped to a specific emotion or variations of the emotion. For example, the emotion "happy" can have variations like "quite happy", "very happy", "less happy". Schröder [11] uses several equations to describe how the prosodic elements vary while changing the emotional states. The parameters are distinguished as: a) "Standard" global parameters: "pitch", "range", "speech rate" and "volume", b) "Non-standard" global parameters: "pitch-dynamics" and "range-dynamics" and c) specific entities like ToBI accents and boundaries.

The MARY (Modular Architecture for Research on speech sYnthesis) Text-to-Speech system [12] is an open-source, Java implemented platform. The system follows the Client-Server (CS) model. Server side executes text preprocessing / normalization, natural language processing, calculation of acoustic parameters and speech synthesis. The client sends to server the requests, including the text to be processed and the parameters for the text handling by the server side. The system is multi-threaded, due to CS implementation, flexible (modular architecture) and XML-based, adding support for DOM and XSLT [12]. MARY TtS includes a number of tools, in order to easily add a new language and build Unit Selection and HMM-based synthetic voices [13-14].

In the present study, we first introduce a novel experimental approach towards the evaluation of emotional prosodic models in Expressive Speech Synthesis. Then we investigate how emotions are communicated through speech/prosodic channel excluding any emotion from content's semantics. The results can be incorporated into TtS systems for the acoustic rendition of document's typographic signals. Our study also builds on the universal character of emotions hypothesis according to which emotion is similarly expressed across languages. It ultimately aims to evaluate this hypothesis by testing an existing, language-agnostic / universal prosodic model for conveying emotions through TtS. In contrast, previous approaches (e.g. [15-17]) to developing a prosodic model for the Greek language are based on the creation of an emotional speech database, along with feature extraction and analysis.

2 The Experimental Approach

We adopted the Self-Assessment Manikin Test (SAM) [18] which measures the emotional response, based on the dimensional approach of emotions [19-20]. SAM test provides to evaluators the ability to avoid the verbal expression of emotions in the assessment. It introduces a quick and easy procedure. This tool has been designed to replace the course of self-assessment of the emotions. In the context of the current

study, we have designed and developed a web-based version of the SAM experimental procedure, similar to the one described in [21].

2.1 Stimulus Design

In order to eliminate any expression of emotions through the verbal/semantic channel, the selection of meaningless pseudo-sentences that resemble normal speech (in both Greek and English TtS) is mandatory. In the Greek version of synthesized stimuli, the selection of the pseudo-sentences was done according to the methodology proposed in [22]. A set of random short sentences was selected and phonemes of the content words of the sentences were replaced, such that pseudo-words were formed. Vowels were replaced by vowels and consonants by consonants, and the syllabic structure and stress of the original word were maintained.

The pseudo-sentences were converted into synthetic speech, using MARY TtS [12]. A number of languages are currently supported by MARY TtS, such as English, German, Russian, Italian, Turkish and Telugu. But the Greek language is not included. Following the basic and necessary steps for the baseline support of a new language in the MARY TtS framework [13] we have developed a HMM-based Greek voice. The dimensional description of emotional states and the prosodic model [11] [23] for *pitch*, *rate* and *volume* has been applied, in order to acoustically render "Pleasure", "Arousal" and "Dominance".

Participants first hear each pseudo-sentence two times and then they assess the emotional state they perceive using manikins. The stimulus presentation sequence (frame) is presented in Fig. 1. The audio cue begins with a short pause (2 seconds), followed by the synthesized pseudo-sentence, a short pause, repetition of the pseudo-sentence and an ending pause.

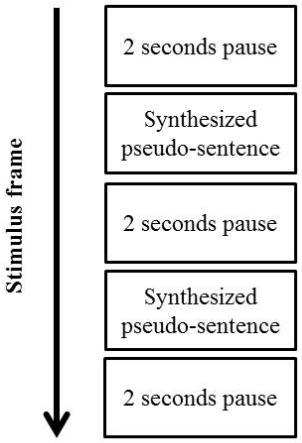

Fig. 1. The stimulus audio sequence (frame) during the experimental procedure

2.2 Stimuli Implementation

We selected the dimensional representation (Table 1) of 10 emotions provided by [24] and the "neutral" emotional state. Fig. 2 presents the scatter graph of the ten emotional states on the "Pleasure - Arousal" grid. Finally, we selected two sentences that can be optimally processed by the TtS system. The sentence "Ηταν όμως λίγο απότομος" was transformed into "Ηταν ένως τίγο αφόθονος" and the second pseudo-sentence was "Η αφιστροπή της σάστες είναι γογεμός". For the English stimuli we selected two pseudo-sentences "Hat sundig pron you venzy" and "Fee got laish jankill gosterr" [9] [25]. The total stimuli were 44 (we implemented 11 emotional states X 2 pseudo-sentences for each version of the TtS language).

Table 1. The 10 emotional states and their corresponding values on the "Pleasure", "Arousal" and "Dominance" dimensions in scale [-1, 1]

Emotional State	Pleasure	Arousal	Dominance
A	-0.51	0.59	0.25
B	-0.60	0.35	0.11
C	-0.64	0.60	-0.43
D	0.81	0.51	0.46
E	-0.63	-0.27	-0.33
F	0.40	0.67	-0.13
G	0.74	-0.13	0.03
H	0.87	0.54	-0.18
I	0.68	-0.46	0.06
J	-0.65	-0.62	-0.33

2.3 Participants and Procedure

In total 15 students (9 male and 6 female) participated in the experiment with an average age of 24.4 year-old (SD=3.5). They were graduate and post-graduate students of the Department of Informatics, University of Athens. Their native language was Greek (with excellent or good proficiency in English). They did not report any hearing problem, and they had none or a little familiarization with synthetic speech or Text-to-Speech systems.

The participants were asked to hear and estimate the emotional state that is communicated by the synthesized speech. They were introduced in the experimental procedure and they had to complete a form about their demographic data (age, occupation, educational level, frequency of computer usage) and that they consent to participate in the experimental test. Then, they were familiarized with the experiment by participating in a short demo session. They had to assess 4 demo stimuli (2 in Greek and 2 in English version). They could repeat the demo session if they believed that they were not familiarized with the procedure. After the demo version of the experiment, they participated in the main procedure, where they assessed the 44 stimuli. The stimuli playback was provided in a random order for each participant, through a high performance headset (AKG K271). The average assessment time was 18.1 minutes (SD=3.3).

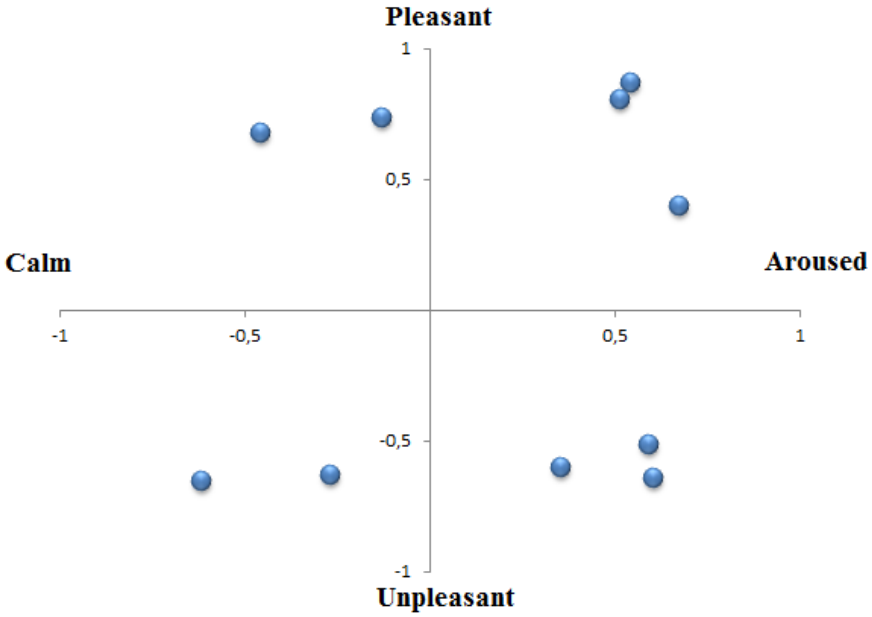

Fig. 2. Scatter graph of the 10 emotional states on the "Pleasure - Arousal" grid

3 Results

Tables 2 and 3 present the results for the Greek pseudo-sentences and Tables 4 and 5 the corresponding English ones. Each table includes the average responses for

Table 2. Average values with their corresponding standard deviation for each emotional state on the Pleasure, Arousal and Dominance dimensions for the Greek pseudo-sentence #1

Greek pseudo-sentence #1						
Emotional State	**Pleasure (SD)**	**Target Value**	**Arousal (SD)**	**Target Value**	**Dominance (SD)**	**Target Value**
A	-0.10 (0.43)	*-0.51*	0.57 (0.26)	*0.59*	0.23 (0.80)	*0.25*
B	-0.17 (0.31)	*-0.60*	-0.33 (0.41)	*0.35*	0.43 (0.50)	*0.11*
C	-0.10 (0.43)	*-0.64*	0.47 (0.30)	*0.60*	0.40 (0.63)	*-0.43*
D	-0.30 (0.41)	*0.81*	0.83 (0.24)	*0.51*	-0.27 (0.62)	*0.46*
E	-0.40 (0.47)	*-0.63*	-0.60 (0.34)	*-0.27*	0.03 (0.52)	*-0.33*
F	0.00 (0.53)	*0.40*	0.90 (0.21)	*0.67*	-0.13 (0.81)	*-0.13*
G	-0.13 (0.23)	*0.74*	-0.20 (0.37)	*-0.13*	0.33 (0.62)	*0.03*
H	0.03 (0.44)	*0.87*	0.67 (0.24)	*0.54*	0.07 (0.59)	*-0.18*
I	-0.10 (0.34)	*0.68*	-0.47 (0.40)	*-0.46*	0.33 (0.72)	*0.06*
J	-0.53 (0.44)	*-0.65*	-0.87 (0.30)	*-0.62*	0.10 (0.71)	*-0.33*
Neutral	0.00 (0.19)	*0.00*	0.00 (0.33)	*0.00*	0.13 (0.61)	*0.00*

Table 3. Average values with their corresponding standard deviation for each emotional state on Pleasure, Arousal and Dominance dimensions for the Greek pseudo-sentence #2

Greek pseudo-sentence #2						
Emotional State	**Pleasure (SD)**	**Target Value**	**Arousal (SD)**	**Target Value**	**Dominance (SD)**	**Target Value**
A	0.07 (0.46)	*-0.51*	0.33 (0.49)	*0.59*	-0.10 (0.57)	*0.25*
B	0.07 (0.37)	*-0.60*	0.10 (0.34)	*0.35*	0.10 (0.57)	*0.11*
C	0.03 (0.48)	*-0.64*	0.43 (0.37)	*0.60*	-0.17 (0.67)	*-0.43*
D	-0.13 (0.55)	*0.81*	0.90 (0.21)	*0.51*	-0.27 (0.80)	*0.46*
E	-0.60 (0.34)	*-0.63*	-0.53 (0.40)	*-0.27*	0.10 (0.60)	*-0.33*
F	0.10 (0.51)	*0.40*	0.77 (0.32)	*0.67*	0.00 (0.76)	*-0.13*
G	-0.20 (0.32)	*0.74*	-0.20 (0.53)	*-0.13*	0.30 (0.49)	*0.03*
H	0.00 (0.57)	*0.87*	0.83 (0.31)	*0.54*	-0.13 (0.88)	*-0.18*
I	-0.17 (0.45)	*0.68*	-0.20 (0.37)	*-0.46*	0.20 (0.41)	*0.06*
J	-0.53 (0.40)	*-0.65*	-0.67 (0.45)	*-0.62*	0.27 (0.56)	*-0.33*
Neutral	-0.07 (0.32)	*0.00*	-0.07 (0.42)	*0.00*	0.23 (0.50)	*0.00*

"Pleasure", "Arousal" and "Dominance" with their standard deviation (in brackets). Also, the initial values of the emotional states used during the speech synthesis procedure for the stimuli implementation is presented (Tables 2-5, named as "Target Value").

We observe that "Arousal" dimension was perceived as intended in both pseudo-sentences (Greek stimuli). The only exception was noticed in Table 2, emotional state B. In the "Pleasure" dimension, the positive valued states were not perceived as intended contrary to negative ones. The "Dominance" dimension did not have the desired results. It is worth noting that for both "Pleasure" and "Arousal' dimensions, neutral state was correctly perceived.

The results for the English pseudo-sentences were slightly worse than those presented for the Greek language. There is a consistency; that "Arousal" dimension was better perceived than "Pleasure" and "Dominance".

Table 4. Average values with their corresponding standard deviation for each emotional state on the Pleasure, Arousal and Dominance dimensions for the English pseudo-sentence #1

English pseudo-sentence #1						
Emotional State	**Pleasure (SD)**	**Target Value**	**Arousal (SD)**	**Target Value**	**Dominance (SD)**	**Target Value**
A	0.10 (0.28)	*-0.51*	-0.07 (0.37)	*0.59*	0.37 (0.58)	*0.25*
B	-0.10 (0.39)	*-0.60*	-0.30 (0.49)	*0.35*	0.23 (0.50)	*0.11*
C	0.20 (0.32)	*-0.64*	0.23 (0.46)	*0.60*	0.33 (0.62)	*-0.43*
D	-0.10 (0.47)	*0.81*	0.67 (0.31)	*0.51*	-0.13 (0.69)	*0.46*
E	-0.53 (0.35)	*-0.63*	-0.67 (0.56)	*-0.27*	0.10 (0.66)	*-0.33*
F	0.00 (0.46)	*0.40*	0.70 (0.37)	*0.67*	-0.20 (0.59)	*-0.13*
G	0.03 (0.35)	*0.74*	-0.33 (0.41)	*-0.13*	0.23 (0.59)	*0.03*
H	0.20 (0.53)	*0.87*	0.57 (0.37)	*0.54*	0.00 (0.82)	*-0.18*
I	0.00 (0.27)	*0.68*	-0.30 (0.49)	*-0.46*	0.13 (0.64)	*0.06*
J	-0.83 (0.31)	*-0.65*	-0.57 (0.56)	*-0.62*	-0.20 (0.80)	*-0.33*
Neutral	-0.10 (0.21)	*0.00*	-0.13 (0.40)	*0.00*	0.20 (0.59)	*0.00*

Table 5. Average values with their corresponding standard deviation for each emotional state on the Pleasure, Arousal and Dominance dimensions for the English pseudo-sentence #2

English pseudo-sentence #2						
Emotional State	**Pleasure (SD)**	**Target Value**	**Arousal (SD)**	**Target Value**	**Dominance (SD)**	**Target Value**
A	-0.10 (0.28)	*-0.51*	-0.03 (0.48)	*0.59*	0.17 (0.52)	*0.25*
B	-0.17 (0.41)	*-0.60*	-0.37 (0.55)	*0.35*	0.27 (0.62)	*0.11*
C	0.07 (0.46)	*-0.64*	0.17 (0.59)	*0.60*	0.23 (0.53)	*-0.43*
D	0.17 (0.56)	*0.81*	0.70 (0.32)	*0.51*	0.00 (0.82)	*0.46*
E	-0.50 (0.46)	*-0.63*	-0.67 (0.36)	*-0.27*	0.03 (0.69)	*-0.33*
F	-0.17 (0.56)	*0.40*	0.53 (0.40)	*0.67*	-0.03 (0.72)	*-0.13*
G	-0.23 (0.37)	*0.74*	-0.37 (0.44)	*-0.13*	0.13 (0.58)	*0.03*
H	0.00 (0.42)	*0.87*	0.67 (0.41)	*0.54*	0.00 (0.73)	*-0.18*
I	-0.33 (0.45)	*0.68*	-0.40 (0.43)	*-0.46*	0.20 (0.41)	*0.06*
J	-0.70 (0.37)	*-0.65*	-0.87 (0.23)	*-0.62*	-0.10 (0.74)	*-0.33*
Neutral	-0.37 (0.40)	*0.00*	-0.30 (0.49)	*0.00*	0.30 (0.59)	*0.00*

4 Conclusions

In this study, we introduce a novel experimental approach towards the evaluation of emotional prosodic models in Expressive Speech Synthesis. This approach is based on the dimensional emotion expressivity and adopts the Self-Assessment Manikin Test. We apply this experimental approach to evaluate the emotional prosodic model, proposed by Schröder [11], for the Greek expressive Text-to-Speech synthesis. We used two pseudo-sentences for each of the Greek and English HMM-based synthetic voices, implemented in the MARY TtS platform. Fifteen native Greek participants were asked to assess eleven emotional states for each sentence.

According to Schröder's results [11] we expected that the best perceived dimension is "Arousal" followed by "Pleasure". In the same study the "Dominance" dimension was not investigated. The preliminary results indicated that "Arousal" dimension was perceived as intended by the participants for both Greek and English pseudo-sentences. "Pleasure" and "Dominance" dimensions were not perceived as accurately as "Arousal". A noticeable result was that Greek pseudo-sentences in the "Pleasure" dimension with positive values were perceived as negative. In contrast, the neutral state for both "Pleasure" and "Arousal" in the Greek version was perceived correctly (especially the first pseudo-sentence).

The proposed experimental approach can also be applied to the study of the degree of speaker's emotional state perception, combining the semantic channel (emotions deriving from text's content) and expressive speech synthesis (ESS). Furthermore, a future implementation would be the adaptation of interaction during the experiment e.g. blind and/or low-vision participants using the haptic modality. This would facilitate to investigate how participants with visual impairment, including blindness, perceive the emotional states through the acoustic channel, using prosodic variations and/or content's semantic information.

Acknowledgements. This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) under the Research Funding Project: "THALIS-University of Macedonia- KAIKOS: Audio and Tactile Access to Knowledge for Individuals with Visual Impairments", MIS 380442.

References

1. Tatham, M., Morton, K.: Expression in Speech: Analysis and Synthesis. Oxford Linguistics, Oxford University Press (2006)
2. Campbell, N., Hamza, W., Hoge, H., Tao, J., Bailly, G.: Editorial Special Section on Expressive Speech Synthesis. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1097–1098 (2006)
3. Kouroupetroglou, G.: Incorporating Typographic, Logical and Layout Knowledge of Documents into Text-to-Speech. In: Encarnacao, P., Azevedo, L., Gelderblom, G.-J., Newell, A., Mathieassen, N.-E. (eds.) Assistive Technology: From Research to Practice, Proceedings of the 12th European AAATE Conference, Vilamoura, Portugal, September 19-22, pp. 708–713. IOS Press (2013), doi:10.3233/978-1-61499-304-9-708
4. Kouroupetroglou, G., Tsonos, D.: Multimodal Accessibility of Documents. In: Advances in Human-Computer Interaction, pp. 451–470. I-Tech Education and Publishing, Vienna (2008)
5. Kouroupetroglou, G., Tsonos, D., Vlahos, E.: DocEmoX: A System for the Typography-Derived Emotional Annotation of Documents. In: Stephanidis, C. (ed.) UAHCI 2009, Part III. LNCS, vol. 5616, pp. 550–558. Springer, Heidelberg (2009)
6. Freitas, D., Kouroupetroglou, G.: Speech Technologies for Blind and Low Vision Persons. Technology and Disability 20, 135–156 (2008)
7. Tsonos, D., Kouroupetroglou, G., Deligiorgi, D.: Regression Modeling of Reader's Emotions Induced by Font Based Text Signals. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2013, Part II. LNCS, vol. 8010, pp. 434–443. Springer, Heidelberg (2013)
8. Abelin, A., Allwood, J.: Cross Linguistic Interpretation of Expressions of Emotions. In: Proceedings of the 8th Simposio Internactional de Communicacion Social, pp. 387–393 (2003)
9. Scherer, K.R., Banse, R., Wallbott, H.G.: Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. Journal of Cross-Cultural Psychology 32(1), 76–92 (2001)
10. Pell, M., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.: Factors in the Recognition of Vocally Expressed Emotions: A comparison of Four Languages. Journal of Phonetics 37(4), 417–435 (2009)
11. Schröder, M.: Expressing degree of activation in synthetic speech. IEEE Transactions on Audio, Speech and Language Processing 14(4), 1128–1136 (2006)
12. Schröder, M., Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology 6, 365–377 (2003)
13. Pammi, S., Charfuelan, M., Schröder, M.: Multilingual Voice Creation Toolkit for the MARY TTS Platform. In: Proceedings of the International Conference on language Resources and Evaluation (LREC), pp. 3750–3756 (2010)

14. Schröder, M., Charfuelan, M., Pammi, S., Steiner, I.: Open source voice creation toolkit for the MARY TTS Platform. In: Proc. of the 12th Conference of the International Speech Communication Association (INTERSPEECH), pp. 3253–3256 (2011)
15. Fakotakis, N.: Corpus Design, Recording and Phonetic Analysis of Greek Emotional Database. In: Proceedings of the International Conference on language Resources and Evaluation (LREC), pp. 1391–1394 (2004)
16. Kostoulas, T., Ganchev, T., Mporas, I., Fakotakis, N.: A real-world emotional speech corpus for modern Greek. In: Proceedings of the International Conference on language Resources and Evaluation (LREC), pp. 2676–2680 (2008)
17. Lazaridis, A., Mporas, I.: Evaluation of Hidden Semi-Markov Models Training Methods for Greek Emotional Text-to-Speech Synthesis. International Journal of Information Technology and Computer Science 05(04), 23–29 (2013)
18. Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25(1), 49–59 (1994)
19. Scherer, K.R.: What are emotions? And how can they be measured? Social Science Information 44(4), 695–729 (2005)
20. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality 11(3), 273–294 (1977)
21. Kouroupetroglou, G., Papatheodorou, N., Tsonos, D.: Design and Development Methodology for the Emotional State Estimation of Verbs. In: Holzinger, A., Ziefle, M., Hitz, M., Debevc, M. (eds.) SouthCHI 2013. LNCS, vol. 7946, pp. 1–15. Springer, Heidelberg (2013)
22. Castro, S., Lima, L., Recognizing, C.F.: emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody. Behavior Research Methods 42(1), 74–81 (2010)
23. OpenMARY, Emotion-to-Mary XSL, `http://mary.dfki.de/lib/emotion-to-mary.xsl/view`
24. James, A., Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of Research in Personality 11(3), 273–294 (1977)
25. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70(3), 614–636 (1996)