Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

The Development of the Athens Emotional States Inventory (AESI): Collection, Validation and Automatic Processing of Emotionally Loaded Sentences

***Theodora Chaspari**, PhD Student*

University of Southern California, Ming Hsieh Department of Electrical Engineering, 3740, McClintock Ave., EEB 400, 90089, Los Angeles, CA, USA

chaspari@usc.edu

Tel: +1-213-7403477

***Constantin Soldatos**, Professor, Director Of Mental Health Care Unit*

Mental Health Care Unit, Evgenidion Hospital, University of Athens

7, Eginitou Str. 11528, Athens, Greece

egslelabath@hol.gr

Tel: +30-210-7250544, Fax: +30-210-7226431

***Petros Maragos**, Professor*

National Technical University of Athens, School of Electrical and Computer Engineering

9, Iroon Polytechniou Str., ECE Building 2.1.24, 15773, Athens, Greece

maragos@cs.ntua.gr

Tel: +30-210-7722360, +30-210-7722420, Fax: +30-210-7723397

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

Abstract

Objectives. The development of ecologically valid procedures for collecting reliable and unbiased emotional data towards computer interfaces with social and affective intelligence targeting patients with mental disorders. *Methods*. Following its development, presented with, the Athens Emotional States Inventory (AESI) proposes the design, recording and validation of an audiovisual database for five emotional states: anger, fear, joy, sadness and neutral. The items of the AESI consist of sentences each having content indicative of the corresponding emotion. Emotional content was assessed through a survey of 40 young participants with a questionnaire following the Latin square design. The emotional sentences that were correctly identified by 85% of the participants were recorded in a soundproof room with microphones and cameras. A preliminary validation of AESI is performed through automatic emotion recognition experiments from speech. *Results.* The resulting database contains 696 recorded utterances in Greek language by 20 native speakers and has a total duration of approximately 28 min. Speech classification results yield accuracy up to 75.15% for automatically recognizing the emotions in AESI. *Conclusions.* These results indicate the usefulness of our approach for collecting emotional data with reliable con- tent, balanced across classes and with reduced environmental variability.

Keywords

expressed Emotion, questionnaires, data collection, factual databases, automatic data processing

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

I. Introduction

The typical concept of computers as tools for information storage, retrieval and management is increasingly being replaced by their view as socially and emotionally intelligent agents that are able to meaningfully interact with humans. Human–computer interaction has long passed the completion of simple procedural tasks and has been redefined by the ability of computer systems to address emotion and affect (Hudlicka 2003; Beale and Peter 2008). This is further confounded by the increasing integration of ambient intelligence environments (De Ruyter et al. 2005) and embodied conversational agents (Nijholt 2003; Foster 2007) to a variety of applications for health, education and commerce, which largely rely on the user's affective responses.

Incorporating affect in human–computer interaction includes a number of steps ranging from sensing, recognizing and modeling the user's state to designing computer systems that adapt to this state and generate appropriate responses (Hudlicka 2003). Their individual study and their integration towards a unified architecture underscores the main research threads of the field. Emotion can be useful in many human–computer applications targeting patients with mental disorders. The quantitative analysis of audiovisual emotional cues could yield objective assessments of patients' emotional expressivity and guide them through assistive intelligent environments and embodied conversational agents (Leite et al. 2013). Quantifying the emotional valence and arousal of individuals with depressive disorders can further provide objective tools to clinicians for making health-related decisions (Moore et al. 2008; Low et al. 2011). The use of biosensors can monitor daily health behaviors (Mitra et al. 2012), identify potential sources of

stressors and capture arousal levels of individuals with developmental disorders (Picard 2009; Leijdekkers et al. 2013).

Towards the automatic recognition of human emotions, appropriate procedures have to be developed to collect ecologically valid data.We present the "Athens Emotional States Inventory" (AESI) in an effort to collect, validate and automatically process items for an emotional database and the correspond- ing acquired data. A debatable issue is the tradeoff between the naturalness of expression and the control of environmental conditions (Douglas-Cowie et al. 2003). The first is related to how "realistically" a sentence is uttered in a given scenario, while the second affects the good quality of recordings. AESI stems from the need to record acted emotional sentences in a controlled environment by preserving as much as possible the naturalness of expression.

During the design of AESI, Greek emotional sentences created by the authors are evaluated based on a survey in terms of their emotional content and recorded in a soundproof room with appropriate technical equipment and procedures. To control for potential differences across emotions and generate a variety of test patches for increased ecological validity, the formation of the survey and the recording procedure follow the Latin square design (Krebs 1999).The same setup is used for recording the sen- tences that were correctly recognized by 85% of the people participating in the survey (seven out of eight sentences per emotion). AESI includes data from five different emotional states (anger, fear, joy, sad- ness and neutral), contains 696 utterances and has overall duration of around 28 min.

We further validate the items of AESI through automatic emotion recognition experiments. We use machine learning techniques to classify the corresponding sentences based on speech-derived prosody, intonation and cepstral features and compare our results with those obtained from largely used emotional databases, such as the Berlin Database of Emotional Speech (EmoDB; Burkhardt et al. 2005). Our classification experiments for AESI reach up to 75.15% unweighted accuracy, which is comparable to the corresponding accuracy of 77.15% achieved for EmoDB with the same setup.

This paper is organized as follows. In Section II, we present previous work specifically related to the design of emotional databases and the methods used for automatic emotion recognition. Details about the procedures followed in AESI in terms of assessing and recording the emotional sentences are described in Sections III and IV, respectively. Section V provides a preliminary analysis of the acquired audio data of AESI in terms of emotion classification performance. A discussion of our work in terms of specific design choices along with related future extensions is given in Section VI and conclusions are provided in Section VII.

II. Related Work

II.1. Design of Emotional Databases

There are three main types of emotional databases, containing acted, natural and artificially elicited emotions. In acted databases professional actors or amateurs are asked to utter sentences or texts conveying specific emotions (Burkhardt et al. 2005; Chiţu et al. 2008; Nwe et al. 2003), to simulate emotional situations (Busso et al. 2008), or to perform a series of emotional facial displays (Kanade et al. 2000). Similar acted data but closer to real-world environments were

collected from movies (Dhall et al. 2012). Natural databases involve real-life situations, such as television shows with emotionally loaded content (Douglas-Cowie et al. 2003; Martin et al. 2009), phone recordings in call centers (Lee and Narayanan 2005) and discussions with various types of emotionally colored semi-automated operators (McKeown et al. 2012). Procedures of artificial induction of mood (Cullen et al. 2006) are designed to induce a specific emotional state of the test subject under controlled conditions. There are also some efforts focusing on the emotional analysis of music (Soleymani et al. 2013).

A detailed survey of speech emotional databases from different languages and with various elicitation methods can be found in (Ververidis and Kotropoulos 2003). The authors point out the extended use of acted and simulated emotions, which provide data from controlled conditions reducing complexity. Taking this into account, AESI contains acted data. Besides recording the predetermined items, this paper also attempts to assess the emotionally loaded sentences with a survey and preserve ecological validity of the recording setup.

II.2. Automatic Emotion Recognition from Speech

Research efforts in automatic emotion recognition have focused on both feature extraction and classification algorithms. Schuller and co-workers have used low-level contours of prosody, voice quality and articulatory information for extracting high-level functionals that describe the emotional content of a sentence (Schuller et al. 2006; Schuller and Rigoll 2006). Vlasenko et al. (2007) have combined diverse time scales of frame- and turn-level to recognize emotions from speech. Other studies have attempted to use the modulation properties of speech signals (Zhou et

al. 1998; Wu et al. 2011) to recognize human affective states. A study surveying the features and classification schemes for speech emotion recognition can be found in (El Ayadi et al. 2011).

In this paper, as a preliminary validation of the data collected from AESI, we use the established prosodic, intonation and cepstral features in emotion recognition research to classify between the various emotional states with Gaussian Mixture Models and Support Vector Machines. The latter consist extensively used machine learning techniques because of their good performance and relative simplicity (Schuller and Rigoll 2006; Vlasenko et al. 2007).

III. AESI Design: Assessment of Emotional Loading of Sentences

AESI contains acted emotions in the effort to produce high-quality audiovisual recordings from a balanced number of classes and subjects. We study four different emotional states (anger, fear, joy, sadness) and the state of neutral emotion, since these are very common in human–computer applications (Section VI). For each emotional state (noted as A, F, J, S and N for anger, fear, joy, sadness and neutral, respectively), there are eight Greek sentences with content indicative of each emotion, resulting in 40 sentences in total. The English translation of these sentences is shown in Table I.

In order to assess the AESI items, we conducted a survey, in which we asked 40 young people to indicate the most appropriate emotional state for every sentence based on its content. For the survey we created eight separate formulations of the questionnaire (Q1–Q8) with a different sequence of the original sentences, ordered according to a Latin square design (Krebs 1999). Specifically, we divided the 40 sentences into eight groups (G1–G8) of five items, each

containing one sentence per emotion. In group 1 (G1), for example, we assigned sentences A1, F1, J1, S1, N1, in group 2 (G2) sentences A2, F2, J2, S2, N2, etc. The order of the sentences within each group was random. At the questionnaire level, each questionnaire Q_i, i=1,..., 8, included the groups of sentences in the following way: G_i, $G_{i\ 1}$, ..., G_{i-2}, G_{i-1}. Therefore in questionnaire Q_1 the groups of sentences were ordered as G_1, G_2, ..., G_8, in Q_2 as G_2, G_3, ..., G_8, G_1, etc. (Figure 1).

The total of 40 people (24 male, 16 female) who participated in the survey were students of the National Technical University of Athens. They were all native Greek speakers with an average age of 23.9 (2.9 SD) years. Although the participants did not have any experience in this kind of task, the questions that were asked can be answered based on common sense and the emotions to be recognized are very distinct from each other.

For each sentence we computed the percentage of people who correctly recognized its corresponding emotional state (Figure 2). In order to ensure the reliability of our database, we only kept the sentences with 85% or higher recognition percentage (i.e., 35 out of 40 people recognized correctly the corresponding emotion). Accordingly, the last sentence among those listed in Table I for each emotional state was not included in the final inventory. The 85% threshold value resulted in removing the ambiguous sentences (such as F8, N8) but also getting balanced emotional classes by preserving the same number of sentences per emotion. This also allowed us to keep the Latin square design formulation simple during the presentation of sentences in the recording procedure (Section 4). Previous studies have used similar techniques

to exclude emotional sentences. In the formation of EmoDB (Burkhardt et al. 2005), for example, naïve listeners were asked to rate the naturalness and emotional tone of already recorded sentences. For the final database, utterances with an emotion recognition rate higher than 80% and naturalness more than 60% were included. Although our task was slightly different in that we preselected the sentences prior to the recordings, we chose the 85% instead of the 80% threshold, because the latter would give us an unbalanced distribution of emotional classes.

IV. Recording Emotional Sentences

In the recordings, we included the 35 sentences (seven per emotion) that were correctly recognized with respect to their emotional content by 85% of the people participating in the survey (Section III). In the following subsections, we describe the equip- ment (Section IV.1) and the recording procedure (Section IV.2).

IV.1. Equipment

To eliminate any echoing, the recording room was soundproof appropriately equipped with double doors and insulated double walls, as well as with enhanced and insulated ceiling and floor. Moreover, acoustic absorption panels were installed on the walls (Vicoustic SuperKit MD55, 30 cm square) (Figure 3).

The acoustic recording equipment included the following commercially available devices, hardware and software:

1. Long gun microphone head Sennheiser ME67 and powering module Sennheiser K6.
2. Eight-channel microphone preamplifier Presonus DigiMax FS.

3. Analog to digital converter with the digital signal handling card TC Electronic Digital Konnekt x32.
4. The software package Studio Production Package Cubase Essential 5 for the computerized handling of the data.

The sampling rate of the recordings was 44.1kHz and the encoding was implemented using 16 bits.

The video equipment consisted of:

1. Two cameras with resolution 720×480 pixels and 6-15mm lens.
2. The MPEG4 digital video recorder XRPLUS XRS 4008 recording at 25 frames/sec.

IV.2. Recording Procedure

Twenty undergraduate university students (12 male, 8 female; age 23.6±1.5 years), all native Greek speakers participated in the recordings. This group of speakers was entirely different from the people that completed the survey (Section III).

During the data collection, same procedure was followed for each person in order to reduce the inherent speaker bias. Every speaker was seated on an armchair in front of a laptop screen in the recording room (Figure 3) and was asked to follow the instructions appearing on the screen. He/she had to express the appropriate emotion corresponding to the spoken sentence. For each of the 35 sentences, four instructions were presented. The speaker had to act according to the corresponding instruction within a predetermined time interval. The instructions and the corresponding waiting time (in parentheses) are given below:

1. In the following sentence you will be asked to embed emotion X (where X stands for anger, fear, joy, sadness, neutral). (waiting time: 10 sec)
2. Memorize the sentence (one of the 35 sentences is shown at the same time). (waiting time: 15 sec)
3. Spell out the previous sentence expressing the corresponding emotion while looking at the camera. (waiting time: 15 sec)
4. Regain neutral emotional state. (waiting time: 10 sec)

The 35 sentences that each person uttered were ordered according to the latin square design (Krebs 1999), as was similarly performed for the 40 sentences of the questionnaire (Section 3).

The resulting database contains 696 utterances (4 out of 700 utterances were omitted by the speakers by accident) and has a total duration of about 28 minutes. Further details can be found in Table 2.

V. Preliminary Validation of AESI through Automatic Emotion Classification

As a first step towards evaluating the feasibility of AESI, we tested its automatic classification accuracy. Traditionally the audio-based emotion recognition uses standard features, such as fundamental phonation, vocal tract features and speech energy, inspired by speech recognition, as reported in (Ververidis and Kotropoulos, 2008). Similar features were used here.

We performed preliminary classification experiments on AESI and EmoDB (Burkhardt, 2005). The latter was selected since it is designed in a similar way to AESI. It contains acted emotional data from 10 professional actors uttering 10 different sentences, each with 7 emotions (anger, fear, joy, sadness, boredom, dislike and neutral). The algorithm used for both datasets

extracts meaningful features from speech, performs feature selection to preserve the most relevant information for our task and classifies among the different emotional states. A schematic representation is shown in Figure 4.

We examined two different sets of features. The first feature set (referred as ``Feature Set 1") captures spectral and prosody information from speech. It contains the first 13 Mel-Frequency Cepstral Coefficients (MFCCs), its first-order derivatives and the fundamental frequency (F0) of speech, resulting in 27 features. We computed the mean and standard deviation of these features over the whole sentence in order to capture the general emotional evolution in speech, resulting in an array of 54 parameters. The second feature set (referred to as ``Feature Set 2") was inspired from the Interspeech 2009 Emotion Challenge. It includes the speech Zero Crossing Rate (ZCR), Root Mean Square (RMS) energy, fundamental frequency (F0), Harmonics-to-Noise Ratio (HNR) and first 12 MFCC coefficients. Based on these 16 low-level descriptors and their first-order derivatives, 12 functionals over each emotional sentence were extracted, including statistical moments, extrema information and linear regression coefficients, resulting in an array of 384 parameters. A detailed description of the second feature extraction scheme, which is well established in the automatic emotion recognition literature (Lee et al., 2011; Wöllmer et al. 2010), can be found in (Schuller et al., 2009).

We performed a series of feature transformation methods in order to compensate for speaker variability and remove any information redundant to our task. First we normalized all samples from the same speaker to have zero-mean and unity standard deviation. We selected the most relevant features according to the Fisher Discriminant Ratio (FDR) criterion (Duda et al., 2000). Setting a cut-off threshold to the FDR value, feature discriminability is increased by

maximizing the inter-class and minimizing the intra-class distance of samples. We note as f_{prc} the percentile of the maximum FDR value from all feature pairs for which the threshold was set. In order to avoid redundant information, we computed the Pearson's correlation coefficients between all pairs of remaining features. If a pair of features had a coefficient greater than a threshold, symbolized as r_{thr}, we removed the feature with the lower FDR. Finally to further increase class separation, we performed Linear Discriminant Analysis (LDA) (Duda et al., 2000).

Our classification scheme includes two different classifiers: the Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs), implemented with HTK (Young et al., 2006) and SVM-light (Joachims et al., 1999), respectively. The classification experiments were performed within a leave-one-speaker-out cross-validation setup, in which sentences from one speaker were included in the test set, sentences from another speaker in the development set, and the remaining were used for training. The development set was used to tune the feature selection and classification parameters and the test served as a blind set to check our results. Cross-validation was performed to avoid issues of data overfit.

During feature selection, we examined the following FDR percentile and Pearson correlation threshold values on the development set: f_{prc}=0.05, … ,0.5 (with step 0.05) and r_{thr}=0.1, … ,0.9 (with step 0.1). The system had to choose between 1 or 2 Gaussian mixtures for the GMM classification and among C={0.01, 0.1, 1, 5, 10} trade-off parameters values for the SVM classifier. All feature transformations were derived based on the train data and then applied on the development and test set at each fold.

The classification results for both feature sets and classifiers are shown in Table 3. Classification accuracy for AESI varies from 49.2% to 75.15%, while for EmoDB results range

between 68.01% and 77.15%. For both datasets, we observe that the second feature set yields better performance, since it has a richer representation of the speech prosody and spectrum. GMMs seem to be more effective in these classification tasks than SVMs. This could be due to the fact that the feature space dimensionality after LDA is fairly low (not exceeding 6 and 4 for EmoDB and AESI respectively) and GMMs might be able to better model such low-dimensional spaces.

Classification accuracy of AESI is comparable to EmoDB for the second feature set, which is the most descriptive. Taking into account AESI's larger speaker and lexical variability, it is reasonable to assume that rich feature representations can yield better performance; that might be the reason why Feature Set 1, which is relatively simple, shows poor accuracy. These results further indicate that the methods proposed in the current study consist a promising framework for designing such databases.

VI. Discussion and Future Extensions

In this paper we presented an acted audiovisual emotional database. We emphasized on collecting high quality audiovisual data of representative affective content, approach which has several benefits. Since all sentences are predetermined with constant linguistic information, the effect of emotion on speech can be easily compared across subjects as well as acoustic features. Visual processing is simplified, because the lighting conditions and the speaker's positioning remain the same during the recordings. No additional tagging is needed, as would happen in a database containing natural flow of speech, because the emotions are predetermined for each sentence. Furthermore, controlling the content of the database results in a balanced distribution

of sentences across emotions, which is not always the case for other types of emotional databases.

The four prototypical emotions of AESI have been the focus of many studies related to mood and anxiety disorders. In fact, as Lara et al. (Lara et al., 2006) have proposed, the relative contribution of anger and fear in different mixtures can distinguish between the different types of mood disorder states and predict their most common episodes. The four emotions of AESI have been studied with respect to the ability of adolescents with mood and anxiety disorders to recognize facial expressions (McClure et al., 2003). In terms of intervention, positive emotions, like happiness, when channeled into appropriate treatment, can be effective in coping with anxiety disorders (Fredrickson, 2000). Of course, other prototypical emotions, such as disgust, surprise, anticipation and trust, can be important to mood and anxiety disorders. The comparison between positive and negative emotions is central to these disorders, as many studies have shown (Watson et al., 1988; Brown et al., 1998; Leppänen, 2006). Therefore, with no purpose of underestimating the importance of the remaining emotions, the existence of both positive and negative emotions in AESI could be helpful towards its clinical applications to mood and anxiety disorders.

The emotions of AESI are heavily studied in the Autism research as well. Hobson et al. (Hobson et al., 1988) examined the way children with Autism match emotional expressions of anger, fear, joy and sadness across different individuals. Yirmiya et al. (Yirmiya et al., 1992) assessed the ability of this population to discriminate between the emotional states of anger, fear, joy, sadness and proudness based on videotaped stories. McIntosh et al. (McIntosh et al., 2006)

analyzed the mimicry of happy and angry emotional expressions of adolescents and adults with ASD.

As a matter of fact, other studies in clinical psychology, e.g., (Grossman and Tager-Flusberg, 2008), have examined certain additional prototypical emotions, such as disgust and surprise which may be implicated in many disorders. Our work specifically addresses a set of emotions commonly referred in mood/anxiety and ASD studies, therefore anger, fear, joy and sadness were studied here. In its current form, AESI could be used to record emotional sentences expressed by patients with mental disorders, which can be automatically analyzed in terms of emotional expressivity (Section V). Nevertheless, in the future, we plan to expand AESI to more emotional states following the same procedure here as described in Sections II and III.

Quantifying the acoustic characteristics of emotions with computerized techniques could find clinical applications to standardized assessments of patients' emotional expressivity. Similar automated analysis of acoustic features has been used to evaluate prosody production of children with ASD (van Santen et al. 2010; Bone et al. 2012; Albornoz et al., 2013; Schuller et al. 2013), patients with depression (Mundt et al., 2007; Moore et al., 2008; Low et al., 2011), Parkinson's disease (Asgari and Shafran, 2010) and other psychiatric disorders (Cohen and Elvevåg, 2014). Because of its predetermined and well-structured setup, AESI could be used in order to collect data from such populations. Analyzing these data could provide functional measures -not easily derived by human experts- that can offer objective evaluation criteria and might be able to stratify large heterogeneous clinical populations into smaller well-defined risk groups. For this reason, we plan to expand AESI to study emotional expression of clinical populations.

One limitation of our study is the use of acted data which might not always simulate the affect found in natural conversational settings. The inherent trade-off between acted and natural emotional databases (Douglas-Cowie, 2003) led to the use of a questionnaire for assessing sentences in terms of their emotional content. In order to minimize variability across people, the latin square design was followed both in the survey and the recordings. Although the natural expression of acted emotional data is not always guaranteed, carefully planned databases, such as AESI, could potentially bridge this gap and be further used for adapting affective speech in personalized HCI applications (Skelley et al., 2006; Sanchez et al., 2010). As previously discussed, such setups can also benefit clinical scenarios in order to minimize variability across the populations that are being compared.

Another limitation is that items of AESI are originally formed and collected in the Greek language. It is actually of the few systematic efforts to record emotional sentences in greek with high quality audiovisual equipment with previous studies focusing on isolated emotional words (Lazaridis et al., 2010) or data from a more unstructured setup (Ververidis et al., 2008). Besides the usefulness of adding more data in the sparse Greek resources, the present study describes methods that are generalizable, since the original AESI items (Table 1) with valid emotional content for most Indo-European cultures can be translated into any other language. The survey and recording procedure (Section IV) can be easily replicated as well.

In this study we provided a preliminary acoustic analysis of AESI with similar emotion classification performance to the widely used EmoDB (Burkhardt et al, 2005). In addition to the standard features stemming from the linear source-filter speech model, we plan to use some non-linear features of the modulation type as in (Dimitriadis et al., 2005) that could potentially better

capture changes in the voice tone, musicality and speech amplitude associated with emotion. Combining those with appropriate information from the visual stream could further benefit our results. In our ongoing work on vision-based emotion recognition, we use face representations and related features from Active Appearance Models (Cootes et al., 2001) that represent the shape and texture of a face as deviations from a predefined exemplar. This could be used for tracking changes of important face parts during affective expressions, such as cheeks, lips and eye corners.

VII. Conclusions

The Athens Emotional States Inventory (AESI) was created in an effort to design ecologically valid acted emotional databases with emphasis on the emotional content and the naturalness of recorded expression, meaning how ``realistic'' is the acted performance. A survey procedure was performed for rating predefined sentences according to their corresponding emotional content. The sentences that were recognized correctly by most of the people participating in the survey were further recorded according to a predefined procedure with constant environmental conditions and same equipment. In order to reduce bias across people, the items in the survey questionnaire as well as in the recording procedure were sorted according to the latin square design. A preliminary validation of the collected data through emotion classification experiments with performance similar to other emotional databases indicates that they contain representative content and can be further used in other affective studies.

In our future work, we plan to expand AESI items to more emotional states, such as stress and anticipation, also found in typical human-computer interaction scenarios. We will also

collect similar data from clinical populations in order to study typical and atypical facets of emotional expressions. Finally, in the context of automatic emotion classification, we are going to study non-linear acoustic features that could better capture the fine emotional speech modulations and combine those with visual cues for acquiring complementary information.

Acknowledgments

The authors would like to thank Professor Thomas Paparrigopoulos, psychiatrist, Dr. Golfo Liamaki, clinical psychologist, for their constructive comments in the early stage of the formation of this paper as well as Professor Athanassios Protopapas, cognitive scientist, for his valuable advice on the recording room and equipment setting.The authors would also like to thank Dr. Stelios Krasanakis, psychiatrist and drama therapist, for his critical help in formulating the AESI sentences. Finally, special thanks are due to Dr. Dimitrios Dimitriadis for his contribution on the classification experiments. P. Maragos' research work was supported in part by the project COGNIMUSE which is implemented under the ARISTEIA Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.

Statement of Interest

None to declare.

References

Albornoz EM, Vignolo LD, Martinez CE, Milone DH. 2013. Genetic wrapper approach for automatic diagnosis of speech disorders related to Autism. Ins: IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI), pp. 387–392.

Asgari M, Shafran I. 2010. Predicting severity of Parkinson's disease from speech. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**,** pp. 5201–5204.

Beale R, Peter C. 2008. The role of affect and emotion in HCI. In: Affect and Emotion in Human-Computer Interaction. Berlin: Springer. p 1–11.

Bone D, Lee CC, Black MP, Williams ME, Lee S, Levitt P, Narayanan S. 2014. The Psychologist as an Interlocutor in Autism Spectrum Disorder Assessment: insights from a study of spontaneous prosody. J Speech Lang Hear Res 57:1162–1177.

Brown TA, Chorpita BF, Barlow DH. 1998. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. J Abnorm Psychol 107(2):179–192.

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. 2005. A database of German emotional speech. In: Proceedings of Interspeech, pp. 1517–1520.

Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, et al. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. J Lang Resour Eval 42(1):335–359.

Chiţu AG, van Vulpen M, Takapoui P, Rothkrantz LJ. 2008. Building a Dutch multimodal corpus for emotion recognition. In: Proceedings of Language Resources and Evaluation Conference (LREC).Workshop on Corpora for Research on Emotion and Affect, pp. 53–56.

Cohen AS, Elvevåg B. 2014. Automated computerized analysis of speech in psychiatric disorders. Curr Opin Psychiatry, 27(3):203–209.

Cootes TF, Edwards GJ, Taylor CJ. 2001. Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(1):681–685.

Cullen C, Vaughan B, Kousidis S, Wang Y, McDonnell C, Campbell D. 2006. Generation of high quality audio natural emotional speech corpus using task based mood induction. In: Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies Extremadura.

De Ruyter B, Saini P, Markopoulos P, Van Breemen A. 2005. Assessing the effects of building social intelligence in a robotic interface for the home. Interact Comput 17(1): 522–541.

Dimitriadis D, Maragos P, Potamianos A. 2005. Robust AM-FM features for speech recognition. IEEE Sig Proc Lett 12(1): 621–624.

Dhall A, Goecke R, Lucey S, Gedeon T. 2012. Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia 19(3):34–41.

Douglas-Cowie E, Campbell N, Cowie R, Roach P. 2003. Emotional speech: towards a new generation of databases. Speech Commun 40(1):33–60.

Duda RO, Hart PE, Stork DG. 2000. Pattern classification. 2nd ed. Oxford: John Wiley & Sons.

El Ayadi M, Kamel MS, Karray F. 2011. Survey on speech emotion recognition: features, classification schemes, and databases. Patt Recog 44(1):572–587.

Foster ME. 2007. Enhancing human-computer interaction with embodied conversational agents. In: Stephanidis C, editor. Proceedings of the Fourth International Conference on Universal Access in Human-Computer Interaction (UAHCI). Ambient Interaction. Berlin: Springer, pp. 828–837.

Fredrickson BL. 2000. Cultivating positive emotions to optimize health and well-being. Prevention & Treatment 3, Article 0001a.

Grossman RB, Tager-Flusberg H. 2008. Reading faces for information about words and emotions in adolescents with autism. Res Autism Spectr Disord 2(1):681–695.

Hobson RP, Ouston J, Lee A. 1988. What's in a face? The case of autism. Br J Psychol 79(4):441–453.

Hudlicka E. 2003. To feel or not to feel: the role of affect in human-computer interaction. Int J Hum Comput Stud 59(1):1–32.

Joachims T, Schölkopf B, Burges C, Smola A. 1999. Making large-scale SVM learning practical. Advances in kernel methods support vector learning. Boston, MA: MIT Press.

Canade T, Cohn JF, Tian Y. 2000. Comprehensive database for facial expression analysis. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53.

Krebs CJ. 1999. Ecological methodology. 2nd ed. Menlo Park, CA: Addison-Wesley Educational Publishers.

Lara DR, Pinto O, Akiskal K, Akiskal HS. 2006. Toward an integrative model of the spectrum of mood, behavioral and personality disorders based on fear and anger traits: I. Clinical implications. J Affect Disord 94(1):67–87.

Lazaridis A, Bourna V, Fakotakis N. 2010. Comparative evaluation of phone duration models for Greek emotional speech. J Comput Sci 6(1):341–349.

Lee CC, Mower E, Busso C, Lee S, Narayanan S. 2011. Emotion recognition using a hierarchical binary decision tree approach. Speech Commun 53(1):1162–1171.

Lee CM, Narayanan SS. 2005.Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Proc 13(1):293–303.

Leijdekkers P, Gay V, Wong F. 2013. CaptureMyEmotion: a mobile app to improve emotion learning for autistic children using sensors. In: Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS), pp. 381–384.

Leite I, Henriques R, Martinho C, Paiva A. 2013. Sensors in the wild: exploring electrodermal activity in child-robot interaction. In: Proceedings of ACM/IEEE International Conference on Human-Robot Interaction, pp. 41–48.

Leppänen JM. 2006. Emotional information processing in mood disorders: a review of behavioral and neuroimaging findings, Curr Opin Psychiatry 19(1):34–39.

Low LS, Maddage MC, Lech M, Sheeber LB, Allen NB. 2011. Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans Biomed Eng 58(3):574–586.

Martin JC, Caridakis G, Devillers L, Karpouzis K, Abrilian S. 2009. Manual annotation and automatic image processing of multimodal emotional behaviors: validating the annotation of TV interviews. Pers Ubiquit Comput 13(1):69–76.

McClure EB, Pope K, Hoberman AJ, Pine DS, Leibenluft E. 2003. Facial expression recognition in adolescents with mood and anxiety disorders. Am J Psychiatry 160(6):1172–1174.

McIntosh DN, Reichmann-Decker A, Winkielman P, Wilbarger JL. 2006. When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. Dev Sci 9(1): 295–302.

McKeown G, Valstar M, Cowie R, Pantic M, Schroder M. 2012. The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans Affect Comput 3(1):5–17.

Mitra U, Emken BA, Lee S, Li M, Rozgic V, Thatte G, et al. 2012. KNOWME: a case study in wireless body area sensor network design. IEEE Commun Mag 50(1):116–125.

Moore E, Clements MA, Peifer JW, Weisser L. 2008. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Trans Biomed Eng 55(1):96–107.

Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. 2007.Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguist 20(1):50–64.

Nijholt A. 2003. Disappearing computers, social actors and embodied agents. In: Proceedings of IEEE International Conference on Cyberworlds, pp. 128–134.

Nwe TL, Foo SW, De Silva LC. 2003. Speech emotion recognition using hidden Markov models. Speech Commun 41(1):603–623.

Picard RW. 2009. Future affective technology for autism and emotion communication. Phil Trans R Soc 364(1):3575–3584.

Sanchez MH, Tür G, Ferrer L, Hakkani-Tür D. 2010. Domain adaptation and compensation for emotion detection. In: Proceedings of Interspeech, pp. 2874–2877.

Schuller B, Reiter S, Rigoll G. 2006. Evolutionary feature generation in speech emotion recognition. In: Proceedings of International Conference on Multimedia and Expo (ICME), pp. 5–8.

Schuller B, Rigoll G. 2006.Timing levels in segment-based speech emotion recognition. In: Proceedings of Interspeech, pp. 1818–1821.

Schuller B, Steidl S, Batliner A. 2009.The INTERSPEECH 2009 Emotion Challenge. In: Proceedings of Interspeech, pp. 312– 315.

Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of Interspeech.

Skelley J, Fischer R, Sarma A, Heisele B. 2006. Recognizing expressions in a new database containing played and natural expressions. In: Proceedings of IEEE International Conference on Pattern Recognition (ICPR), pp. 1220–1225.

Soleymani M, Caro MN, Schmidt EM, Sha CY, Yang YH. 2013. 1000 songs for emotional analysis of music. Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM), pp. 1–6.

van Santen JPH, Prud'hommeaux E, Black LM, Mitchell M. 2010. Computational prosodic markers for autism. Autism 14(3):215–236.

Ververidis D, Kotropoulos C. 2003. A state of the art review on emotional speech databases. In: Proceedings of Richmedia Conference, pp. 109–119.

Ververidis D, Kotropoulos C. 2006. Emotional speech recognition: Resources, features, and methods. Speech Commun 48(1):1162–1181.

Ververidis D, Kotsia I, Kotropoulos C, Pitas I. 2008. Multi-modal emotion-related data collection within a virtual earthquake emulator. In: Proceedings of Language Resources and Evaluation Conference (LREC), Morocco.

Vlasenko B, Schuller B, Wendemuth A, Rigoll G. 2007. Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. In: Proceedings of Interspeech, pp. 2249–2252.

Watson D, Clark LA, Carey G. 1988. Positive and negative affectivity and their relation to anxiety and depressive disorders. J Abnorm Psychol 97(3):346–353.

Wöllmer M, Schuller B, Eyben F, Rigoll G. 2010. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE J Selected Topics Sig Proc 4(1):867–881.

Wu S, Falk T, Chan W. 2011. Automatic speech emotion recognition using modulation spectral features. Speech Commun 53(1):768–785.

Yirmiya N, Sigman MD, Kasari C, Mundy P. 1992. Empathy and cognition in high-functioning autism. Child Dev 63(1):150–160. Young SJ, Evermann G, Gales M, Hain T, Kershaw D, Liu X, et al. 2006. The HTK Book. Cambridge, England: Entropic Cambridge Research Laboratory.

Zhou G, Hansen J, Kaiser J. 1998. Classification of speech under stress based on features derived from the nonlinear Teager energy operator. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 549–552.

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

Table 1: Sentences with emotional content.

Anger	A1) Stop this music immediately, or I will call the police.
	A2) How dare you touch my staff.
	A3) I am very irritated by your inconsistency.
	A4) Leave my room at once.
	A5) Don't you yell at me.
	A6) Your behavior is unacceptable.
	A7) The clothes are messy again.
	A8) You are always late.*
Fear	F1) There is a strange noise behind the door.
	F2) Please don't hit me.
	F3) The car goes so fast, that I cannot control it.
	F4) Do you think the police will stop me for speeding.
	F5) I am so afraid to get into the sea, which looks too deep.
	F6) This dog seems so wild, I am sure he bites.
	F7) I doubt if I can find a new job at this high unemployment.
	F8) I won't have this vaccine with unknown side-effects.*
Joy	J1) These are great news.
	J2) Your present is awesome.
	J3) The grades are announced and I got an 'A'.
	J4) At last. I am off for holidays tomorrow.
	J5) I won the bet.
	J6) This was a great victory.
	J7) Fortunately I will visit my family's house next weekend.
	J8) Congratulations for your success.*
	S1) My beloved dog died.
	S2) Though my grandmother was old, I will still miss her.

Sadness	S3) It is really a pity that this disaster happened. S4) Unfortunately his heart was too weak. S5) It is a shame that breaking up is the only solution. S6) Our experiments led to meaningless results. S7) I failed my exams, although I was sure I had passed them. S8) I am so sorry that we cannot change his mind.*
Neutral	N1) Aluminum frames are more resistant than wooden ones. N2) Today I will go to the supermarket; tomorrow is a holiday. N3) The rust gradually destroys the iron objects. N4) All the dust needs to be removed from the books. N5) An electrician is more prosperous than a plumber. N6) Whole wheat bread is healthier than processed bread. N7) The house was closed for a long time and needs cleaning. N8) Using new generation bulbs is energy saving.*

Each emotional state (anger, fear, joy, sadness, neutral) contains eight sentences with content representative of the corresponding emotion.
* not included in final version of AESI (as explained in section III)

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

Table 2: Characteristics of AESI database in terms of speakers, utterances and duration.

Total number of speakers					20			
Number of sentences per speaker					35			
Number of repetitions per sentence					1			
Total number of utterances					696			
Total duration of the recordings					27 min, 51 sec			
Number of utterances per emotion								
Anger (A)					139			
Fear (F)					139			
Joy (J)					139			
Sadness (S)					140			
Neutral (N)					139			
Mean duration per sentence (sec)								
A1	3.14	F1	1.75	J1	1.31	S1	1.89	N1	3.40
A2	1.84	F2	1.60	J2	2.61	S2	2.78	N2	2.75
A3	2.06	F3	3.39	J3	1.77	S3	1.90	N3	2.26
A4	1.99	F4	3.03	J4	2.04	S4	2.03	N4	1.75
A5	1.80	F5	3.29	J5	1.20	S5	2.22	N5	4.01
A6	1.88	F6	3.63	J6	1.64	S6	2.72	N6	2.44
A7	1.79	F7	3.12	J7	3.25	S7	2.75	N7	3.00

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

Table 3: Classification accuracy (%) of emotions for EmoDB and AESI.

Dataset	Classifier	Feature Set 1 [a]	Feature Set 2 [b]
AESI	GMM	60.15	75.15
	SVM	49.20	68.01
EmoDB	GMM	71.30	77.15
	SVM	72.93	75.27

Experiments are performed with two classifiers (GMM, SVM) and on two different feature sets.

[a] Sentence-level statisticals of the Mel Cepstral Coefficients (MFCCs) and pitch.

[b] Sentence-level descriptors of the speech Zero Crossing Rate (ZCR), Root Mean Square (RMS) energy, fundamental frequency (F_0), Harmonics to Noise Ratio (HNR) and MFCCs.

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

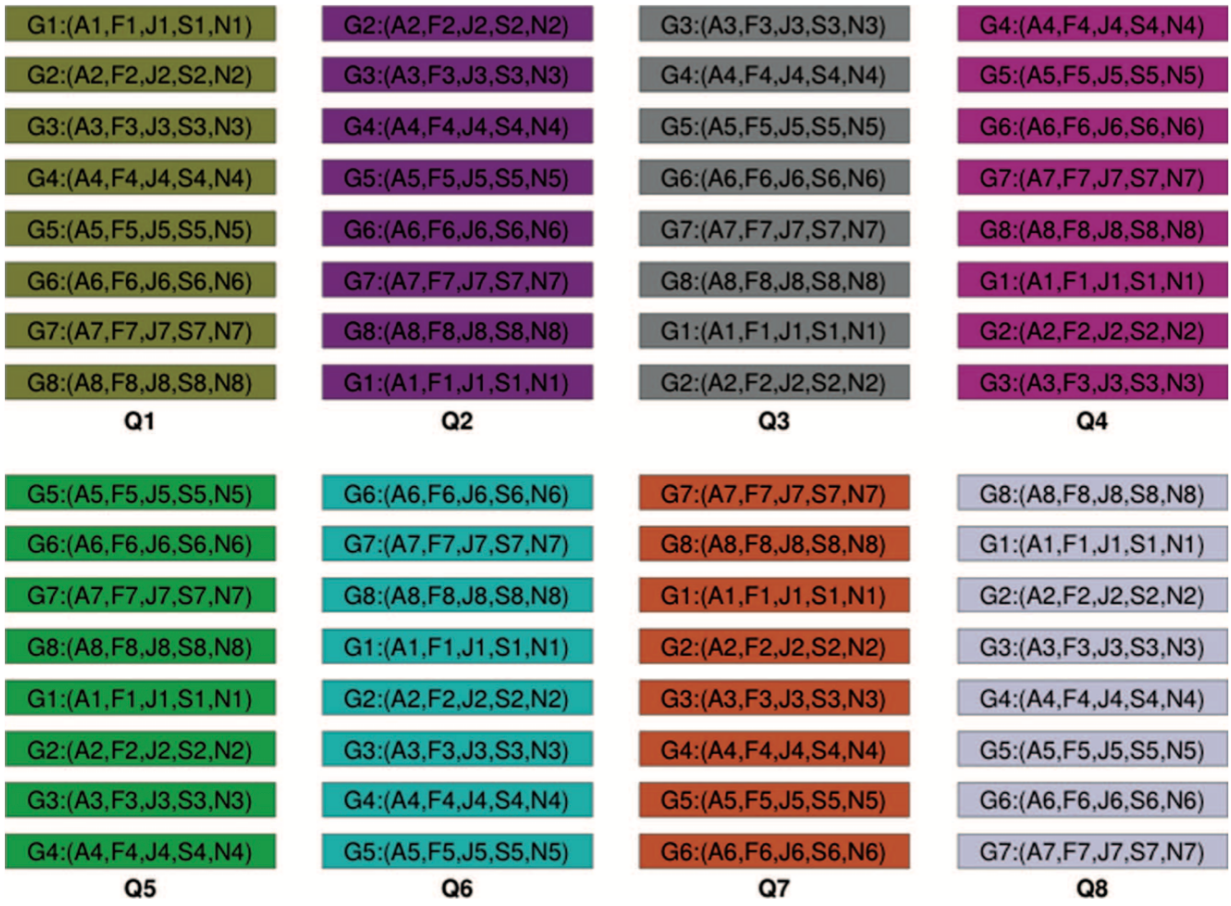

Figure 1: Latin square design for internal differentiation of the order of the AESI items. (Q1-Q8: questionnaires, G1-G8: groups of 5 sentences, A1-A8: anger, F1-F8: fear, J1-J8: joy, S1-S8: sadness, N1-N8: neutral). Each questionnaire contains the eight groups of five sentences in different order. Five sentences (one per emotion) sorted randomly are included in every group.

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

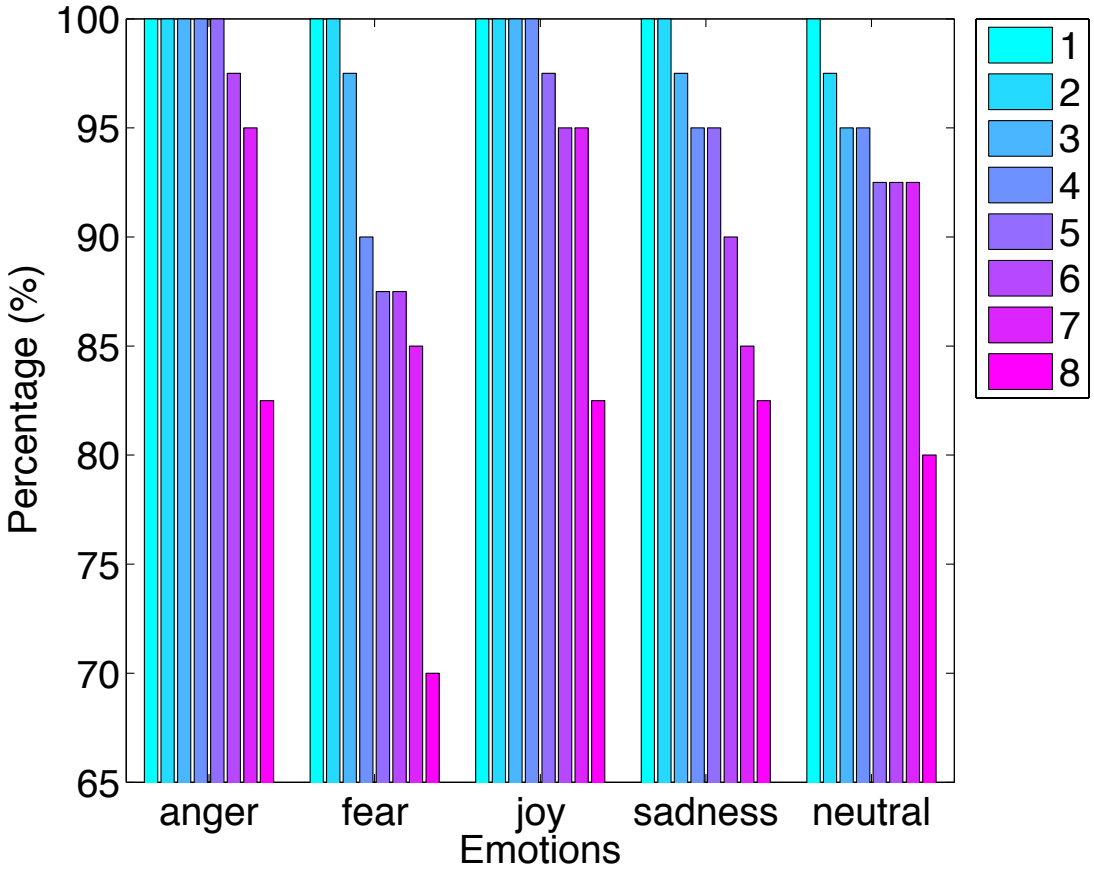

Figure 2: Human recognition percentage (%) of emotional states (anger, fear, joy, sadness, neutral) for each sentence (1-8).

Please cite this paper as: Theodora Chaspari, Constantin Soldatos & Petros Maragos. 2015. The development of the Athens Emotional States Inventory (AESI): collection, validation and automatic processing of emotionally loaded sentences, The World Journal of Biological Psychiatry 16(5):312-322.

Figure 3: Recording room setting. Each person was seated on the armchair in front of the computer screen, which provided the stimuli for uttering the emotional sentences. The two microphones and the absorption panels on the walls are shown.

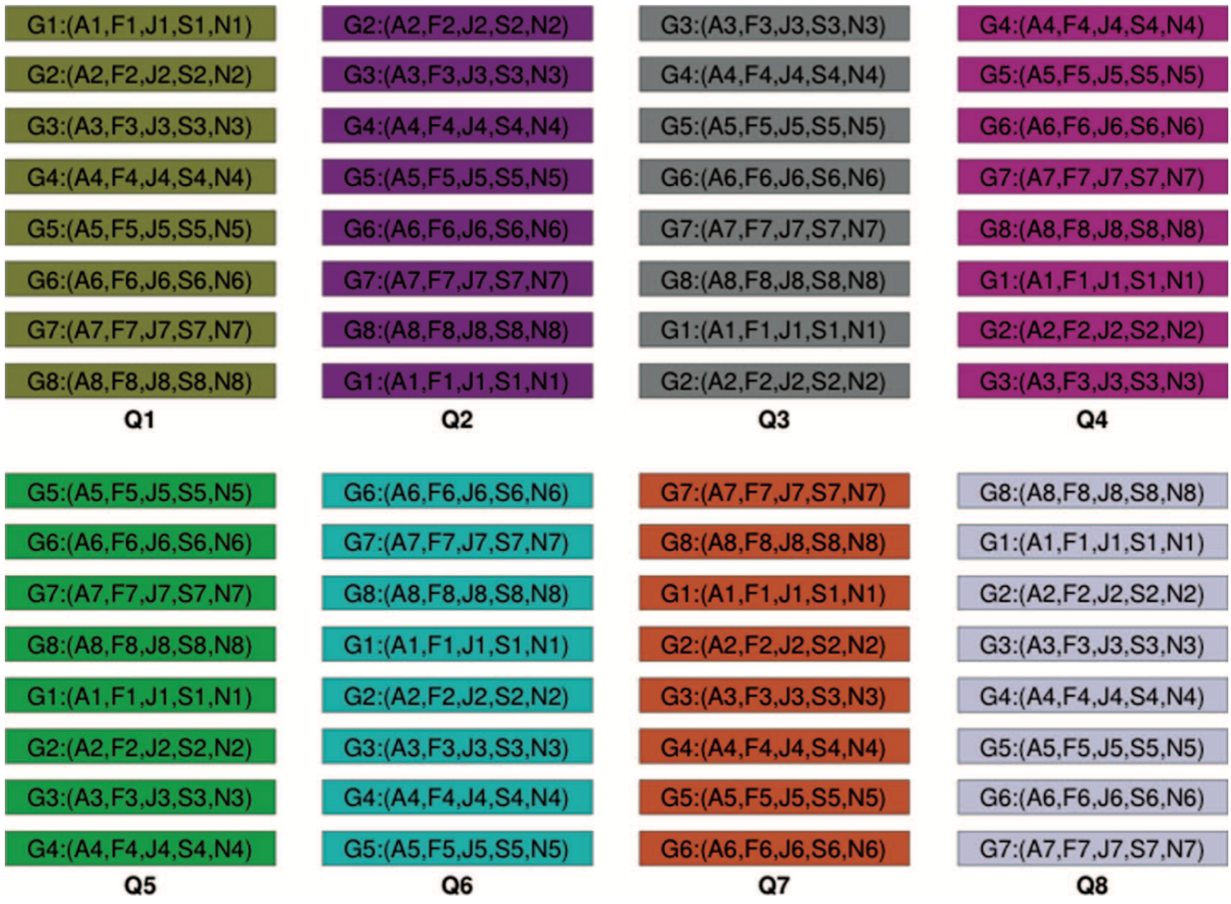

Figure 4: Feature extraction, feature selection and classification scheme. The speech signal is transformed into low-level descriptors of the prosody and the spectrum. High-level functionals are extracted based on each emotional sentence. Feature selection was performed to omit the redundant information and increase class discriminability. The resulting feature vector was fed to the classifier to take the final decision of emotion for the sentence.