Expert Systems with Applications 39 (2012) 11072–11079

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Affective speech interface in serious games for supporting therapy of mental disorders

Theodoros Kostoulas [a,*], Iosif Mporas [a], Otilia Kocsis [a], Todor Ganchev [a], Nikos Katsaounos [a], Juan J. Santamaria [b], Susana Jimenez-Murcia [b], Fernando Fernandez-Aranda [b], Nikos Fakotakis [b]

[a] Wire Communications Laboratory, Department of Electrical and Computer Engineering, University of Patras, 26500 Rion-Patras, Greece
[b] Department of Psychiatry, University Hospital of Bellvitge-IDIBELL and Ciber Fisiopatologia Obesidad y Nutricion (CIBEROBN), 08907 Barcelona, Spain

ARTICLE INFO

Keywords:
Speech interface
Serious games
Mental disorders
Emotion recognition
Affect recognition
Speech recognition

ABSTRACT

We describe a novel design, implementation and evaluation of a speech interface, as part of a platform for the development of *serious games*. The speech interface consists of the speech recognition component and the emotion recognition from speech component. The speech interface relies on a platform designed and implemented to support the development of *serious games*, which supports cognitive-based treatment of patients with mental disorders. The implementation of the speech interface is based on the Olympus/RavenClaw framework. This framework has been extended for the needs of the specific serious games and the respective application domain, by integrating new components, such as emotion recognition from speech. The evaluation of the speech interface utilized purposely collected domain-specific dataset. The speech recognition experiments show that emotional speech moderately affects the performance of the speech interface. Furthermore, the emotion detectors demonstrated satisfying performance for the emotion states of interest, *Anger* and *Boredom*, and contributed towards successful modelling of the patient's emotion status. The performance achieved for speech recognition and for the detection of the emotional states of interest was satisfactory. Recent evaluation of the serious games showed that the patients started to show new coping styles with negative emotions in normal stress life situations.

1. Introduction

Over the past decade, the market growth of the games industry increased the demand of user-friendly interfaces, among which are multimodal game-oriented interfaces that involve spoken interaction. This also holds for the so called *serious games*, or persuasive computer and video games, used as educational tools, or as mean for presenting or promoting a certain point of view (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005; Beale, Kato, Marin-Bowling, Guthrie, & Cole, 2007; Bergeron, 2008). Serious games can be considered as a kind of entertainment-education, which is designed to provide and engage self-reinforcing context, by motivating and educating players towards non-game events or processes (Kalapanidas et al., 2009). Recently, serious games were proved useful therapeutics tools, in support of traditional methods, for treatment of mental disorders (Fernandez-Aranda et al., 2012; Jiménez-Murcia et al., 2009; Santamaria et al., 2011).

The use of new technologies has been applied for a range of mental illnesses (Botella et al., 2004; Griffiths, 2004), reflecting the possible benefits of some videogames (Griffiths, 2004). For example, Russoniello, O'Brien, and Parks (2009) found that playing video games can increase mood and decrease stress. During the last years well-produced serious games are informative and might lead to potential behavioural improvements for patients suffering from a range of medical illnesses (Santamaria et al., 2011). However, current studies assessing the effectiveness of these games have some limitations. Some methodological shortcomings are present, and most of the studies did not control for influencing several factors, such as anxiety, depression or other psychological factors. In the majority of the serious games, psychopathological features of the patients were rarely considered and finally, there is a lack of studies on how serious games affect adult clinical populations. Therefore, additional controlled research is needed to define the specific psychological mechanisms involved in the improvement processes of patients and to learn how a game-based approach affects health behaviours.

In this study, we investigate the design, implementation and evaluation of the purposely designed speech interface of the PlayMancer[1] platform, used within the serious mini-games, in support of the therapy of mental disorders. This platform allows augmenting

* Corresponding author. Tel.: +30 2610 969808; fax: +30 2610 997336.
E-mail address: tkost@upatras.gr (T. Kostoulas).

[1] European Commission co-funded project, which supported the implementation of a platform for serious games.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.03.067

the gaming experience with innovative modes of interaction between the player and the game world (Jiménez-Murcia et al., 2009; Kocsis, Ganchev, Mporas, Papadopoulos, & Fakotakis, 2009), but also the evolvement of the principles for "Universally Accessible Games" into action-based 3D games. The speech-based interface, as presented in this study, is part of a multi-modal and multi-sensor interface developed in the context of the PlayMancer project and is implemented utilizing the Olympus/RavenClaw framework (Bohus & Rudnicky, 2003). The performance of the speech and emotion recognition components is measured on purposely collected, domain-specific datasets, in Spanish language.

This work is organised as follows: Section 2 describes the methodology followed towards designing, implementing and evaluating the speech interface, namely the design of the serious games and the serious games platform, and the dataset – participants used in the evaluation of the speech components. Section 3 details the results of the evaluation and Section 4 offers discussion on the achievements of this work and concluding remarks.

2. Methods

The methodology followed for the design, implementation and evaluation of the speech interface consists of the following milestones, which we detail in the present section:

- design of the serious games for supporting cognitive behavioural and addictive disorders therapy
- design of the serious games platform
- design and implementation of the speech interface:
  (i) speech recognition component
  (ii) emotion recognition component
- design and implementation of the evaluation dataset.

2.1. The serious games

Literature and available clinical evidence suggest that, despite individual differences, people suffering from mental disorders share a host of vulnerabilities, cognitive processes and risk factors that inform clinical recommendations for screening, prevention, diagnostic and treatment. While many approaches for the treatment of these disorders are promoted, the most widely accepted, and most well supported by clinical evidence, are those based on Cognitive-Behavioural Therapy (CBT). CBT is the evidence-based treatment of choice of several mental disorders. CBT is based on the idea that several factors are interacting together: cognitions (how we think), emotions (how we feel) and behaviour (how we act) (van Bastelaar, Pouwer, Cuijpers, Twisk, & Snoek, 2008).

The design of the PlayMancer mini-games was based on a cognitive-behavioural model (van Bastelaar et al., 2008), which usually takes into account attitudinal, behavioural, cognitive and emotional processes in individuals who are interacting with the environment. Up until now, in order to achieve simple behaviours, cognitive models have successfully been applied by designing specific video-games (Brezinka, 2008). However, in our case, the purpose of knowing the attitudinal and motivation of individual's actions was crucial. Therefore, when selecting and designing the specific scenario, specific personality, attitudinal and motivational processes involved in Pathological Gambling and Eating Disorders, based on our previous conducted research, were taken into account. Basically, the main parameters considered in our studies included: shared personality traits (Fernández-Aranda et al., 2006; Álvarez-Moya et al., 2007), emotional variables (Jiménez-Murcia et al., 2007) and behavioural and cognitive processes (Fernández-Aranda et al., 2009; Jiménez-Murcia et al., 2007). At the same time, we were trying to design a game where the player's interest was maintained (fun and exciting), while focused on the main therapeutic goals.

The PlayMancer mini-games are designed as a complementary tool in the treatment of chronic mental disorders (mainly eating disorders and behavioural addictions). Shared biological, emotional and attitudinal vulnerabilities are being described among different mental disorders (Álvarez-Moya et al., 2007). Further shared vulnerabilities have been also described in impulse control disorders and eating disorders. In particular, bulimia nervosa and pathological gambling have been hypothesized to be associated with dysfunction in the brain's reward system (Vicentic & Jones, 2007). Such underlying processes, might explain the reinforcing efficacy of both gambling and binge eating and purging. Insofar, as personality profiles are hypothesized to reflect underlying neurotransmitter function (Ribasés et al., 2008), they may be a valuable window into the nature of this dysfunction and a means to identify similarities and differences between the two disorders both in men and women.

The PlayMancer mini-games introduce the player to an interactive scenario, where the final goal is to increase his general problem solving strategies, self-control skills and control over general impulsive behaviours. After using the games, specific targeted attitudinal, emotional and behavioural changes are expected by the subject. Each task will permit access to one or several types of resources, which will facilitate and improve the game character's, and hence the player's relaxation techniques and planning skills. The games will encourage the player to learn and develop new confrontation strategies (Fernandez-Aranda et al., 2012).

The PlayMancer mini-games designed to support CBT of patients with behavioural and addictive disorders, rely on the speech interface, on the reliable detection of emotions/affective states during the game play (Jiménez-Murcia et al., 2009; Kocsis et al., 2009). To this end, much research has been conducted in the area of affect/emotion recognition from speech (Batliner, Steidl, Hacker, & Nöth, 2008; Cowie et al., 2001; Schuller, Batliner, Steidl, & Seppi, 2011; Schuller et al., 2010). The speech and emotion recognition components of the PlayMancer multimodal interface provide real-time feedback to the game interaction manager, so that the game-play is adjusted to the user condition in each particular moment. For example, if dysfunctional emotions are detected (e.g. *Anger)*, the game becomes more difficult, in order to increase the emotion awareness capacity of the patients, but also to reinforce them to learn additional strategies to gain self-control, following a negative reinforcement process.

2.2. The serious game platform

The design of the serious games platform is based on the Service Oriented Computing paradigm (Papazoglou, Traverso, Dustdar, & Leymann, 2007). The generic architecture of the serious gaming platform is illustrated in Fig. 1. Different service blocks are developed and interconnected to compose the overall serious games environment. The modular architecture is flexible and the system is able to operate with subsets of the service blocks. Further, integration of new services is permitted and allows smooth interoperation of the system.

The central part of the serious games platform is the Unity Game authoring tool (Unity, 2010), which has been enhanced with a number of new interfaces and components, namely:

- Semantic fusion component: Provides support and adaptation of the multimodal input of the serious games platform.
- Game interaction manager: Defines the rules and actions of the game itself. It provides to the game developer the means for implementing the core of the game scenario.

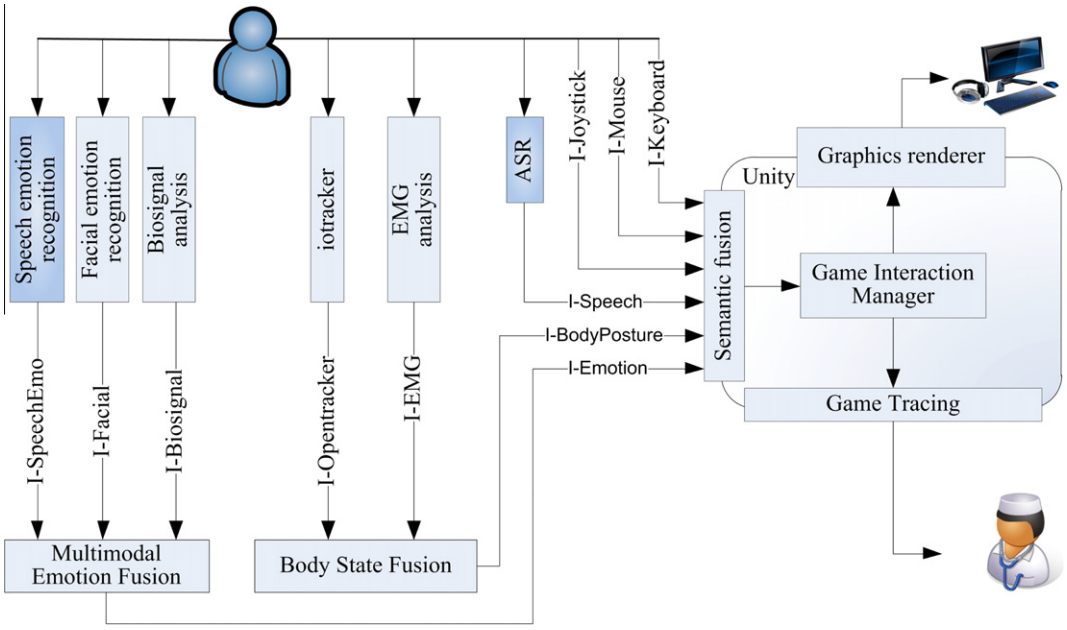

Fig. 1. Block diagram of the PlayMancer platform.

- Game Tracing: In order to allow the medical personnel to review at a later time the evolution of the game and identify the reasons behind a reaction of the patients, the game tracing component available with Unity has been enhanced to record also the multimodal input and emotions identified by the serious games platform.

Depending on the game needs with respect to the recognition of the emotional state or body position, different combinations of multimodal inputs are used. The multimodal emotion fusion and the body state fusion components are responsible to combine and correlate the different inputs and provide a standardized input to Unity (I-Emotion and I-BodyPosture, respectively). The modular architecture of the platform offers simplicity in adding new input devices and in linking them together, i.e., the only update needed is in the multimodal emotion and body state fusion components. The body state fusion component utilizes the input the input from the EMG (Electromyography) and motion tracking, in order to provide a better model of the patients activity and the under training body parts.

The multimodal emotion fusion component can receive input from the *speech*, emotion and biosignal emotion recognition components, when available. It resolves the outputs of the individual emotion detectors using a rule-based schema and provides a robust estimation of the player's emotion. The ASR component (speech recognition and understanding) recognizes a set of spoken commands in Spanish language, as these are determined by the user requirements (D2.4, 2010).

The outputs of the speech interface (speech and emotion recognition components) are processed by the Semantic fusion component, the output of which serves as input for the Game Interaction Manager which defines the rules and actions of the mini-games (D3.3, 2011). This is done either directly (in the case of the speech recognition component) or indirectly (through the multimodal emotion fusion component).

2.3. The speech interface

The speech interface is implemented using the Olympus/RavenClaw framework. The Olympus/RavenClaw framework (Olympus) (Bohus & Rudnicky, 2003) is an open-source flexible speech environment, created at Carnegie Mellon University, mainly designed to help researchers in the implementation of conversational agents, so they can test their ideas without having to build a system from scratch. Olympus framework consists of individual components in the form of executable binaries that communicate to each-other by using socket technology based on the TCP/IP network protocol.

Originally, Olympus was designed for building spoken dialogue systems based on the RavenClaw (Bohus & Rudnicky, 2003) dialogue engine. Selected Olympus framework components were qualified in order to make them act as a *speech interface* to third-party software platforms that support some kind of dialogue management. Also, the qualified Olympus-based speech interface was enhanced with the Speech-based Emotion Recognition Component.

Audio server, voice activity detection, speech recognition, natural language understanding, and speech-based Emotion Recognition components constitute the core of the speech interface for serious games. In detail, the audio server captures audio signals from an audio source, e.g., a microphone. It then sends these processed signal segments to one or more decoding engines (speech recognition / emotion recognition). The decoding engines act as decoding servers for the audio server, returning the decoded hypotheses of the speech signal. One hypothesis per server is qualified as the final result and disposed as an xml-formatted string to specific TCP/IP socket.

The speech interface consists of the speech recognition component and the emotion recognition component, which are detailed below.

2.3.1. Speech recognition component

A general structure of an HMM-based system is illustrated in Fig. 2. During the training phase, speech data (after being subject to pre-processing and speech parameterization) and their corresponding transcriptions are utilized for building the acoustic model. The pre-processing consists of (i) frame blocking every 10 ms with window length 25.6 ms, (ii) energy-based voice activity detection on frame level and (iii) separation of the speech segments from the rest audio input. Low-pass and pre-emphasis filtering is applied on each extracted speech segment and the

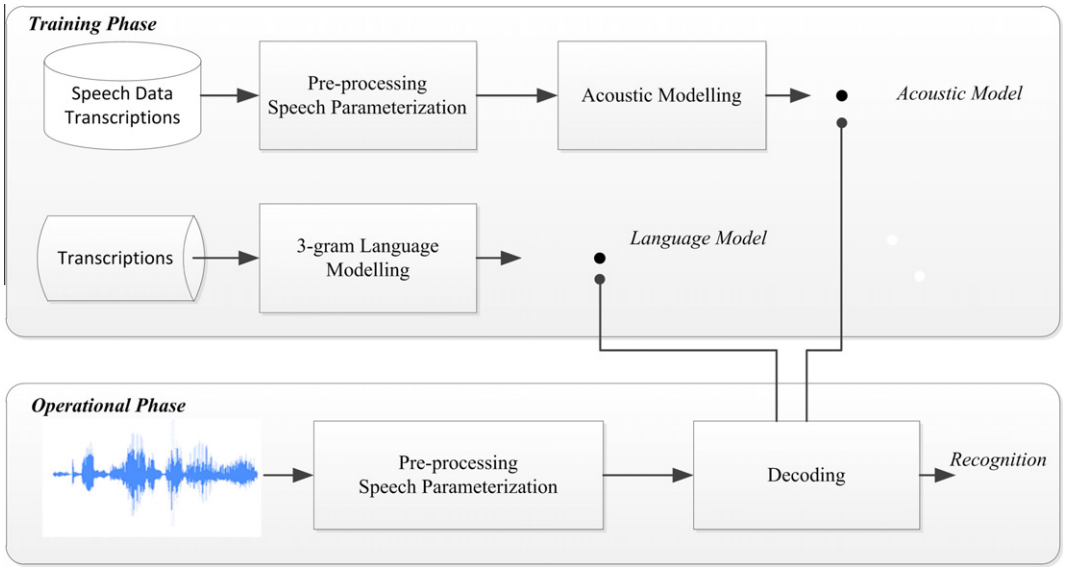

Fig. 2. Block diagram of the HMM-based speech recognition procedure.

corresponding speech frames are multiplied with a Hamming window of the same length. For each windowed frame a parametric vector is computed and consists of the 12 first Mel frequency cepstral coefficients together with the 0th coefficient. For each speech segment cepstral mean normalization is applied. The sequence of static parametric vectors is expanded with their delta and double delta coefficients, resulting to 39-dimention feature vectors, which are used as input to the speech decoder.

The language model consists of 3-gram word models, where the 3-gram probabilities were computed corresponding transcriptions of the serious games scenarios. For the construction of the language model we exploited the CMU Language Model Toolkit (Clarkson and Rosenfeld, 1997).

In the operational phase, the input speech is subject to the same pre-processing and speech parameterization. The resulting feature vector is compared against an acoustic model, consisting of one HMM for every context-independent and context-dependent sound unit. The resulting acoustic score is combined with the score of a language model by the decoder. The language score consists of the probability of each word of a vocabulary to appear after a preceded word sequence. The sequence of words with the highest overall score is the recognized output.

During decoding the input speech waveforms to word transcriptions, the default parameter values of the CMU Sphinx III recognizer were used. The language model weight was set equal to 9.5. For the decoding of the speech waveforms we relied on the open source CMU Sphinx III engine (Sphinx3, 2011). We utilized a general purpose acoustic model trained on the Spanish SpeechDat(II) telephone-speech database (Moreno, 1997) and an application-dependent language model. The acoustic model consists of one 3-state HMM for each of the Spanish phones of the SAMPA alphabet (Wells, 1997). Each HMM state was modelled by a mixture of 8 continuous Gaussian distributions. The state distributions were trained with parametric vectors, which were produced using the pre-processing procedure described above. The decoder utilized context-dependent phone models, while all HMM states were tied to 5000 senones.

2.3.2. Emotion recognition component

Initial specifications detailed by the user requirements defined a wide range of emotion states that need to be controlled, such as anger, boredom, joy, neutral, sadness, and surprise (D2.4, 2010). A further evolution to the mini games scenarios forced reducing the number of emotions that are used in the game to only three, namely *Anger*, *Boredom* and *Neutral* (D2.4, 2010).

The architecture of the emotion detector built for the PlayMancer games, illustrated in Fig. 3, is motivated by previous experience in designing and implementing emotion/affect recognition components (Kostoulas, Ganchev, & Fakotakis, 2010). Both the training and the operational phase make use of the same speech same pre-processing and parameterization, which are based on the openSMILE speech parameterization (Eyben, Wollmer, & Schuller, 2009). The speech parameters extracted stick to state of the art recent findings (Schuller, Steidl, & Batliner, 2009): Sixteen low-level descriptors are extracted (zero crossing rate, root mean square frame energy, pitch frequency, harmonics to noise ratio, 12 Mel-frequency cepstral coefficients) and their delta coefficients are computed. Then, 12 statistic functions are applied on sentence level (mean, std, range, kurtosis, skewness, minimum and maximum value and relative position, two linear regression coefficients and their mean square error) (Schuller et al., 2009) in order to obtain the final feature vector.

The training phase utilizes unlabelled speech data towards creating the universal background model (UBM). This model is assumed to represent the general properties of speech (Dempster, Laird, & Rubin, 1977; Reynolds & Rose, 1995). The emotional data are used to create the respective emotion models by maximum a-posteriori adaption (MAP) (adaptation of the means only) of the UBM (Reynolds, Quatieri, & Dunn, 2000).

The operational/testing phase corresponds to computing the log-likelihoods for the input data belonging to each of the emotion models. Depending on the game design a decision threshold can be applied for concluding to a recognized emotion or the log-likelihood can be directly fed to the appropriate component (e.g. multimodal emotion fusion component) for further processing.

2.4. Evaluation dataset – participants

The collected data consists of recordings from 18 healthy people: eight males and ten females. The choice of healthy subjects for the creation of the speech corpora was imposed for ethical reasons, as those are defined by the ethical committee. The age reported by the test subjects was in the range between 18 and 43 years old, with mean value of 29.3.

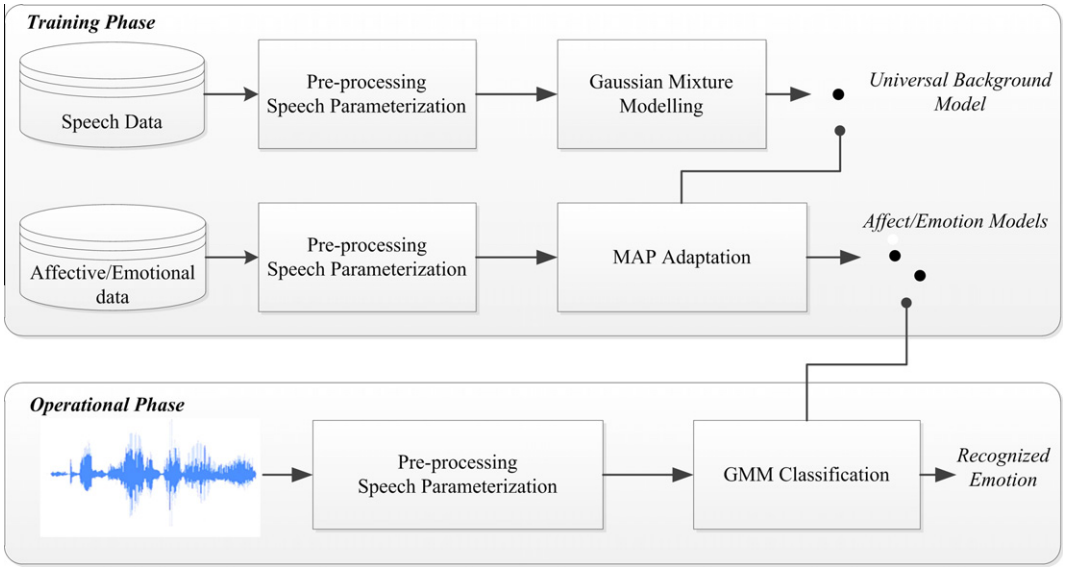

Fig. 3. Block diagram of the emotion detector.

Table 1
Example scenario-related questions and answers for RS002.

Scenario no.	Scenario description	Example question	Example answer
SC001	Making the puzzle	Do you think you could do it, if I gave you a time bonus?	Answer NEUTRAL: How much additional time will you give me?
SC002	Diving free	Your first assignment is to learn how to dive. Are you ready for it?	Answer JOY: Great! I love diving!
SC003	Fishing	Maybe you could try to catch some fishes to put in the aquarium?	Answer SADNESS: I don't know... It sounds too difficult for me...
SC004	Taking Photos	Do you see the animals in front of you? You can take a picture if you want.	Answer BOREDOM: Taking pictures is not very interesting...
SC005	Gardening	Your garden is empty. Don't forget you can find seeds on the island and plant them.	Answer SURPRISE: Oh right! I completely forgot to look for seeds!
SC006	Sailing	Be careful! The sail is about to break!	Answer ANGER: I know! I know! It just doesn't work!
SC007	Climbing	If you want to discover new islands, you have to climb this mountain.	Answer SADNESS: As always, I will fall... Again and again... I'll never find any new island...
SC008	Generic encouragement/ warning messages	You seem a bit tense. Do you want to start the relaxation task?	Answer NEUTRAL: Good idea. Thank you.

Table 2
Speech recognition results.

Test session	Language model	WRR (%)
RS001	RS001	88.3
RS001	RS001 + RS002	84.3
RS002	RS002	71.9
RS002	RS001 + RS002	69.3

The speech corpus utilized in the evaluation of the PlayMancer speech interface consists of two sessions, which are part of the PlayMancer multimodal affect database. This database was purposely designed and implemented in support of research and development activities in the serious game environment (Kostoulas et al., 2010) for the mental disorders application domain.

During the first recording session (RS001), no interaction scenarios were envisaged. The participants uttered prompts presented to them through the graphical user interface. The content of the utterances was related to the vocabulary expected within the user's scenarios, according to specifications of the PlayMancer game scenario.

Table 3
Equal error rate of the emotion detectors for models of different complexity (bold values correspond to the lowest equal error rates achieved).

Number of mixtures	Emotional state	
	Anger (%)	Boredom (%)
1	18.18	**25.00**
2	15.15	**25.00**
4	**14.39**	33.71
8	16.67	31.06
16	20.45	36.36
32	27.27	37.12
64	24.24	47.35
128	47.73	52.27
256	59.47	45.45

In the second recording session (RS002), the participant answered spontaneously to scenario-specific questions provided. Specifically, the participants were asked to act upon requested emotions (anger, boredom, joy, neutral, sadness, and surprise). Each scenario was inspired from the design of the serious games,

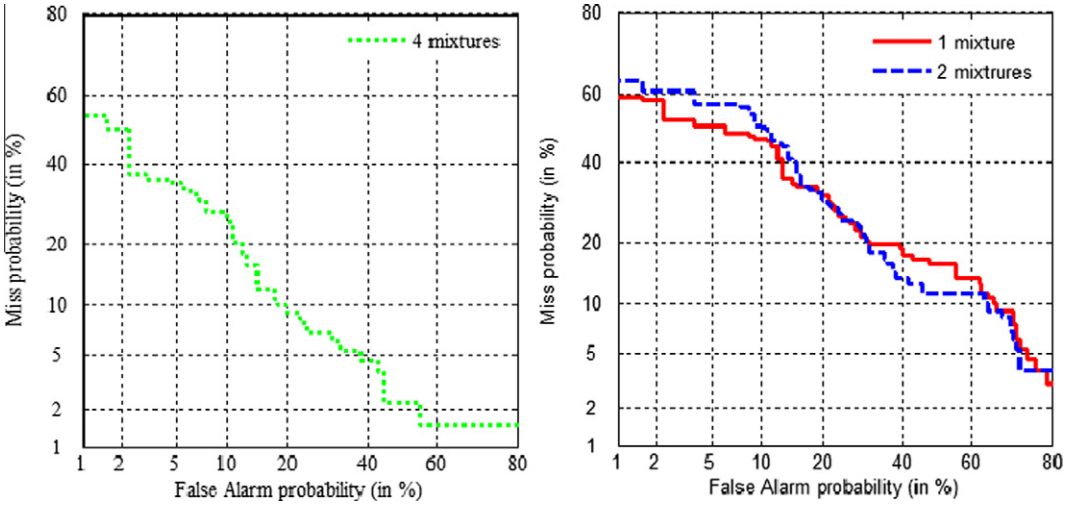

Fig. 4. DET curves for the emotion detectors for the optimal results obtained: anger detector (green dotted line), boredom detector (blue dashed line, red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

as those were detailed by the user requirements. Example scenarios and respective questions and answers are detailed in Table 1.[2]

For each scenario, suggested answers were presented; though each participant was encouraged to use her/his own words and expressions to answer. For both recording sessions, the speech signal was captured using a conventional lapel microphone with sampling rate of 44.1 kHz, single channel, resolution 16 bits. All audio files were stored in .WAV format.

In the evaluation performed here, down-sampled version of the recordings 8 kHz, single channel, with resolution of 16 bits per sample was used. In the speech recognition experiments we made use of the data collected during the first and second recording sessions. In the emotion recognition experiments we used the data of the second recording session, RS002. All speech data from one speaker were discarded due to a failure when validating her data. The remaining data were split in training/testing sets (11 subjects for training, 6 for testing) with respect to age, gender, place of origin and accent distributions.

3. Results

In this section we detail the results for the validation of the speech and emotion recognition components of the PlayMancer speech interface. The experimentations were performed on the training/testing splits described in Section 2.4.

3.1. Speech recognition component

In order to evaluate the performance of the speech interface we performed several experiments on different evaluation data, experimental conditions and configuration settings of the speech recognition engine. The performance of the speech interface was measured in terms of percentage of the word recognition rate (WRR).

The speech recognition accuracy was measured utilizing language models trained with the corresponding transcription of each interaction session, as well as utilizing a more general language model trained with the transcriptions of both sessions, i.e. RS001+RS002. The achieved recognition results are shown in Table 2.

As can be seen in Table 2, the prompted speech together with the RS001-trained language model offered the highest word recognition rates, approximately 88.3%, followed by the more general language model, RS001+ RS002, which achieved accuracy of approximately 84.3%.

In the second recording session the performance was lower than the prompted speech by more than 15%. In detail, for the case of domain-dependent language model the performance dropped to approximately 71.9%, while for the general language model case, RS001+RS002, the performance was approximately, 69.3%. This drop in the performance of the speech recognition component for the case of RS002 is owed to the emotional input speech and use of acoustic models built from neutral speech only.

In both cases the recognition accuracy could be improved with the employment of task grammars for the game scenarios, instead of the use of language models.

3.2. Emotion recognition component

In order to identify the optimal performance of the detectors, we experimented with different number of Gaussian components in the mixtures {1, 2, 4, 8, 16, 32, 64, 128, 256}. The maximum number of iterations was set to 1000 and the criterion for terminating the EM algorithm was the error decrement with less than 10^{-5}. Table 3 summarizes the results obtained for the *Anger* and *Boredom* detectors in terms of Equal Error Rate (EER).

The performance of the emotion recognition component is better for the detection of anger comparing to boredom detection, which was expected due to the characteristics of anger speech and the larger distance from the neutral status in the activation-evaluation space (Whissell, 1989). The DET (Detection Error Trade-off) curves for the emotion detectors with the lowest EER in Table 3 are shown in Fig. 4.

The DET curves offer threshold-independent evaluation of the emotion models, allowing inspection of the detector's performance in various working points (e.g. higher false alarm probability and lower miss probability and vice-versa). The working point of the emotion detectors is defined by the threshold, over which a decision about the emotional state of the input utterance is made. This threshold can be static (set prior the interaction of the patients with the mini games), or dynamic (related to the current state of the game, the patient's specific mental illness, etc.) Further research efforts shall aim at adaptation of the emotion models using additional data, obtained from the interaction of healthy subjects

[2] Spanish translation used is available on request, from the corresponding author.

with the mini games. This will enable modelling some game-specific emotional states and improving the accuracy of the emotion detectors.

4. Discussion and conclusion

We investigated the performance of the speech interface of the PlayMancer platform in the context of serious games. The PlayMancer platform is built on top of existing open-source software and the RavenClaw/Olympus dialog management platform, aiming to provide a novel development framework for serious games development. The existing components are augmented to support multimodality, to be adaptable to context changes, to user preferences, and to game tasks.

The design and implementation was motivated from previous research in speech recognition (Mporas, Ganchev, Kocsis, & Fakotakis, 2011a, 2011b; Mporas, Ganchev, Siafarikas, & Kostoulas, 2007; Mporas, Kocsis, Ganchev, & Fakotakis, 2010). The experimental results indicate that the speech recognition performance for emotional speech is reduced moderately, and thus acoustic models built from emotional speech are required for optimal performance. This observation is in line with previous research on the effect of emotional speech in speech recognition (Kostoulas, Mporas, Ganchev, & Fakotakis, 2008). Though, the employment of task grammars can improve significantly the performance of speech recognition, action, which can be considered in featured updates of the serious games.

The development of the emotion recognition from speech component resulted from previous experience on the field (Kostoulas, Ganchev, & Fakotakis, 2008, 2010; Kostoulas, Ganchev, Lazaridis, & Fakotakis, 2011; Kostoulas, Ganchev, Mporas, & Fakotakis, 2008). Furthermore, the performance achieved for the detection of the emotional states of interest is satisfactory (Batliner et al., 2008; Schuller et al., 2011). Though the results reported correspond to experiments performed with speech data gathered from healthy subjects, recent research has shown that the emotion detector can generalize on other data, despite the mild effect of the mismatch between training and operational conditions (Brendel, Zaccarelli, Schuller, & Devillers, 2010; Kostoulas, Ganchev, & Fakotakis, 2010).

Further, recent evaluation of the Playmancer serious games for CBT show that the patients started to show new coping styles with negative emotions in normal stress life situations, additional generalization patterns, and more self-control strategies when confronted with them (Fernandez-Aranda et al., 2012).

Acknowledgements

This work was supported by the PlayMancer project (FP7-ICT-215839-2007), which was co-funded by the Seventh Framework Programme of the European Commission and CIBER (initiative of Instituto Salud Carlos III). The authors wish to thank the European Commission and CIBER, as well as all members of the project consortium for their support.

References

Álvarez-Moya, E. M., Jiménez-Murcia, S., Granero, R., Vallejo, J., Krug, I., Bulik, C. M., et al. (2007). Comparison of personality risk factors in bulimia nervosa and pathological gambling. *Comprehensive Psychiatry, 48*, 452–457.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis a game without guns. *Educational Technology Research and Development, 53*, 86–107.

Batliner, A., Steidl, S., Hacker, C., & Nöth, E. (2008). Private emotions versus social interaction: A data-driven approach towards analysing emotion in speech. *User Modeling and User-Adapted Interaction, 18*, 175–206.

Beale, I. L., Kato, P. M., Marin-Bowling, V. M., Guthrie, N., & Cole, S. W. (2007). Improvement in cancer-related knowledge following use of a psychoeducational video game for adolescents and young adults with cancer. *Journal of Adolescent Health, 41*, 263–270.

Bergeron, B. (2008). Learning & retention in adaptive serious games. *Studies in Health Technology and Informatics, 132*, 26–30.

Bohus, D., & Rudnicky, A. I. (2003). RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Eurospeech 2003* (pp. 597–600).

Botella, C., Villa, H., García, P. A., Quero, S., Baños, R., & Alcaniz, M. (2004). The use of VR in the treatment of panic disorders and agoraphobia. *Studies in Health Technology and Informatics, 99*, 73–90.

Brendel, M., Zaccarelli, R., Schuller, B., & Devillers, L. (2010). Towards measuring similarity between emotional corpora. In *LREC 2010* (pp. 58–64).

Brezinka, V. (2008). Treasure Hunt-a serious game to support psychotherapeutic treatment of children. *Studies in Health Technology and Informatics, 136*, 71–76.

Clarkson, P., & Rosenfeld, R. (1997). Statistical language modeling using the CMU-Cambridge toolkit. In *Eurospeech 2007* (pp. 2707–2710).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE, 18*, 32–80.

D2.4, D. (2010). Refined requirements and specifications. In *PlayMancer project FP7 215839*.

D3.3, D. (2011). Final Playmancer 3D dialogue-enabled game engine prototype. In *PlayMancer project FP7 215839*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological), 39*, 1–38.

Eyben, F., Wollmer, M., & Schuller, B. (2009). OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In *4th International humaine association conference on affective computing and intelligent interaction* (pp. 1–6): IEEE.

Fernández-Aranda, F., Jiménez-Murcia, S., Álvarez-Moya, E. M., Granero, R., Vallejo, J., & Bulik, C. M. (2006). Impulse control disorders in eating disorders: clinical and therapeutic implications. *Comprehensive Psychiatry, 47*, 482–488.

Fernandez-Aranda, F., Jimenez-Murcia, S., Santamaria, J. J., Gunnard, K., Soto, A., Kalapanidas, E., et al. (2012). Video games as a complementary tool in mental disorders: Playmancer a european multicenter study. *Journal of Mental Health*. doi: 10.3109/09638237.2012.664302.

Fernández-Aranda, F., Núñez, A., Martínez, C., Krug, I., Cappozzo, M., Carrard, I., et al. (2009). Internet-based cognitive-behavioral therapy for bulimia nervosa: a controlled study. *CyberPsychology & Behavior, 12*, 37–41.

Griffiths, M. (2004). Can Videogames Be Good for Your Health? *Health Psychology, 9*, 334–339.

Jiménez-Murcia, S., Álvarez-Moya, E. M., Granero, R., Aymami, M. N., Gómez-Peña, M., Jaurrieta, N., et al. (2007). Cognitive-behavioral group treatment for pathological gambling: analysis of effectiveness and predictors of therapy outcome. *Psychotherapy Research, 17*, 544–552.

Jiménez-Murcia, S., Fernández-Aranda, F., Kalapanidas, E., Konstantas, D., Ganchev, T., Kocsis, O., et al. (2009). Playmancer project: A serious videogame as an additional therapy tool for eating and impulse control disorders. *Annual Review of Cyber Therapy and Telemedicine, 7*, 163–166.

Kalapanidas, E., Fernandez-Aranda, F., Jimenez-Murcia, S., Kocsis, O., Ganchev, T., Kaufmann, H., et al. (2009). PlayMancer: Games for health with accessibility in mind. *Communications & Strategies DigiWorld Economic Journal, 73*, 105–119.

Kocsis, O., Ganchev, T., Mporas, I., Papadopoulos, G., & Fakotakis, N. (2009). Multimodal system architecture for serious gaming. *Artificial Intelligence Applications and Innovations III, 296*, 441–447.

Kostoulas, T., Ganchev, T., & Fakotakis, N. (2008). Study on speaker-independent emotion recognition from speech on real-world data. In *Verbal and nonverbal features of human-human and human-machine interaction* (vol. 5042, pp. 235–242).

Kostoulas, T., Ganchev, T., & Fakotakis, N. (2010). Affect recognition in real life scenarios. In *Toward autonomous, adaptive, and context-aware multimodal interfaces. Theoretical and practical issues* (pp. 429–435).

Kostoulas, T., Ganchev, T., Lazaridis, A., & Fakotakis, N. (2011). Enhancing emotion recognition from speech through feature selection. In *TSD 2010* (pp. 338–344): Springer.

Kostoulas, T., Ganchev, T., Mporas, I., & Fakotakis, N. (2008). A real-world emotional speech corpus for modern Greek. In: *LREC 2008* (pp. 2676–2680).

Kostoulas, T., Kocsis, O., Ganchev, T., Santamaría, J. J., Jiménez-Murcia, S., Moussa, M. B., Magnenat-Thalmann, N., & Fakotakis, N. (2010). The PlayMancer database: a multimodal affect database in support of research and development activities in serious game environment. In *LREC 2010* (pp. 3011–3015).

Kostoulas, T., Mporas, I., Ganchev, T., & Fakotakis, N. (2008). The effect of emotional speech on a smart-home application. In *IEA/AIE 2008* (pp. 305–310).

Moreno, A. (1997). SpeechDat Spanish database for fixed telephone network. Corpus Design Technical Report, SpeechDAT Project LE2-4001.

Mporas, I., Ganchev, T., Kocsis, O., & Fakotakis, N. (2011a). Context-adaptive preprocessing scheme for robust speech recognition in fast-varying noise environment. *Signal Processing, 91*, 2101–2111.

Mporas, I., Ganchev, T., Kocsis, O., & Fakotakis, N. (2011b). Dynamic selection of a speech enhancement method for robust speech recognition in moving motorcycle environment. In *36th International conference on acoustics, speech and signal processing* (pp. 5176–5179): IEEE.

Mporas, I., Ganchev, T., Siafarikas, M., & Kostoulas, T. (2007). Comparative evaluation of speech parameterizations for speech recognition. In *19th International conference on tools with artificial intelligence* (vol. 2, pp. 510–513): IEEE.

Mporas, I., Kocsis, O., Ganchev, T., & Fakotakis, N. (2010). Robust speech interaction in motorcycle environment. *Expert Systems with Applications, 37*, 1827–1835.

Papazoglou, M. P., Traverso, P., Dustdar, S., & Leymann, F. (2007). Service-oriented computing: State of the art and research challenges. *Computer, 40*, 38–45.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*, 19–41.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on, 3*, 72–83.

Ribasés, M., Fernández-Aranda, F., Gratacòs, M., Mercader, J. M., Casasnovas, C., Núñez, A., et al. (2008). Contribution of the serotoninergic system to anxious and depressive traits that may be partially responsible for the phenotypical variability of bulimia nervosa. *Journal of Psychiatric Research, 42*, 50–57.

Russoniello, C., O'Brien, K., & Parks, J. (2009). EEG, HRV and psychological correlates while playing bejeweled II: A randomized controlled study. *Studies in Health Technology and Informatics, 144*, 189–192.

Santamaria, J. J., Soto, A., Fernandez-Aranda, F., Krug, I., Forcano, L., Kalapanidas, E., Gunnard, K., Lam, T., Raguin, T., Davarakis, C., Menchon, J. M., & Jimenez-Murcia, S. (2011). Serious Games as additional Psychological Support: A review of the literature. Cyberpsychology and Behaviour, Therapy, 469–476.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication, 53*, 1062–1087.

Schuller, B., Steidl, S., & Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. In *Interspeech 2009* (pp. 312–315).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. (2010). The Interspeech 2010 Paralinguistic Challenge. In *Interspeech 2010* (pp. 2794–2797).

Sphinx3 (2011). *The CMU sphinx group open source speech recognition engines.* Available from cmusphinx.sourceforge.net/ (accessed 09.04.2012).

Unity (2010). *Unity game development tool.* Available from unity3d.com/ (accessed 09.04.2012).

van Bastelaar, K., Pouwer, F., Cuijpers, P., Twisk, J., & Snoek, F. (2008). Web-based cognitive behavioural therapy (W-CBT) for diabetes patients with co-morbid depression: Design of a randomised controlled trial. *BMC Psychiatry, 8*, 9.

Vicentic, A., & Jones, D. C. (2007). The CART (cocaine-and amphetamine-regulated transcript) system in appetite and drug addiction. *Journal of Pharmacology and Experimental Therapeutics, 320*, 499–506.

Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. *Handbook of Standards and Resources for Spoken Language Systems, 4*.

Whissell, C. (1989). The dictionary of affect in language. *Emotion: Theory, Research, and Experience, 4*, 113–131.