FILTWAM and Voice Emotion Recognition

FILTWAM and Voice Emotion Recognition

Kiavash Bahreini[1], Rob Nadolski[1], and Wim Westera[1]
[1] Centre for Learning Sciences and Technologies (CELSTEC), Open University Netherlands
Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands
{kiavash.bahreini, rob.nadolski, wim.westera}@ou.nl

Abstract. This paper introduces the voice emotion recognition part of our framework for improving learning through webcams and microphones (FILTWAM). This framework enables multimodal emotion recognition of learners during game-based learning. The main goal of this study is to validate the use of microphone data for a real-time and adequate interpretation of vocal expressions into emotional states were the software is calibrated with end users. FILTWAM already incorporates a valid face emotion recognition module and is extended with a voice emotion recognition module. This extension aims to provide relevant and timely feedback based upon learner's vocal intonations. The feedback is expected to enhance learner's awareness of his or her own behavior. Six test persons received the same computer-based tasks in which they were requested to mimic specific vocal expressions. Each test person mimicked 82 emotions, which led to a dataset of 492 emotions. All sessions were recorded on video. An overall accuracy of our software based on the requested emotions and the recognized emotions is a pretty good 74.6% for the emotions happy and neutral emotions; but will be improved for the lower values of an extended set of emotions. In contrast with existing software our solution allows to continuously and unobtrusively monitor learners' intonations and convert these intonations into emotional states. This paves the way for enhancing the quality and efficacy of game-based learning by including the learner's emotional states, and links these to pedagogical scaffolding.

Keywords: Game-based learning; human-computer interaction; multimodal emotion recognition, real-time voice emotion recognition, microphone.

1 Introduction

During the last decade, several new technologies have been adopted by e-learning specialists for enhancing the effectiveness, efficiency and attractiveness of e-learning [1, 2, 3]. Nowadays, learners are often used to the web-based delivery of e-learning content and Web 2.0 affordances when communicating, working and learning together with their peers in distributed (a)synchronous settings [4, 5].
Our research aims for online game-based learning of communication skills. Game-based learning has several advantages: 1) it is a didactical approach that seems to be in-line with the learners' interests [6], 2) it is very popular nowadays [6], 3) it seems

very fruitful to people [7], 4) can be very effective for skills training (see e.g., [8]), and 5) motivating [9]. This approach focusses on learning instead of entertainment, can be quite informal and offers a lot of possibilities to the learners based on learners' input (audio and video). We expect that using technology, affective computing tool, and a pure web-based training system might be insufficient to encourage people to improve their communication skills as the training of such skills needs a lot of recurrent practice. For this purpose, FILTWAM will be deployed with a game-based didactical approach and integrated with EMERGO. EMERGO is a methodology and toolkit for the development and delivery of serious games [10].

FILTWAM uses webcams and microphones to interpret the emotional state of people during their interactions with a game-based learning environment for triggering timely feedback based upon learner's facial expressions and verbalizations. It is meant for discerning the following emotions: sadness, anger, disgust, fear, surprise, happiness, and neutral. It basically offers software with a human-machine interface for the real time interpretation of emotion that can be applied in game-based learning (i.e., affective computing tool). The proof of concept study reported here is a follow up of our previous study [11] and aims to extend our software for multi-modal emotion recognition with a voice emotion recognition part. We will only investigate the opportunities of a microphone for gathering affective user data in an online game-based learning context. An affective computing tool is the development of a system, which is able to recognize, interpret, and simulate human emotions. Our affective computing tool is built upon existing research [12, 13, 14, 15, 16, 17, 18]. Linking two modalities (face expression and voice intonation) into a single system for affective computing analysis is not new and has been studied before [19, 20, 21, 22, 23, 24]. A recent review study by [25] shows that the accuracy of detecting one or more basic emotions is greatly improved when both visual and audio information are used in classification, leading to accuracy levels between 72% to 85%.

It is commonly acknowledged that emotions are an important factor in any learning process, since it influences information processing, memory and performance [26]. Also, feedback based on emotional states may enhance the learners' awareness of their own behavior. This seems relevant for communication skills training. Hence, automated emotion detection as explained in this paper may compensate for the limited number of trainers that are available for online training of communication skills in compare to face-to-face situations [27]. Interacting with digital learning environments is important for the learners; however recent developments of input devices (such as microphones) for interacting with these environments are still underexploited. Such devices not only offer opportunities for more natural interactions with the online game-based learning applications, but also offer ways of gathering affective user data during learning.

There is a growing body of studies on online and automatic voice emotion recognition systems [28, 29, 30]. In one well-known study [28] the researchers have presented a framework for classifier creation for both offline and real-time voice emotion recognition with a specific interface for German speakers. They allowed creating a dataset for joy, satisfaction, anger, and frustration emotions in the context of a training procedure for non-expert users. They received a huge range of accuracy from 24% to 74% for their offline speaker dependent classification approach among 29 users. They reported that this big variation is according to the uncontrolled audio

recording situation in which all the recordings prepared by the users at home. They finally reported that the recognition accuracy using the naïve Bayes (NB) classifier in machine learning approach was 41% for 10 native German speakers.

One study in real-time signal processing and recognition [29] presented a dedicated framework to support a tool for developing the social signal interpretation (SSI). The researchers introduced flexible software architecture to manage audio and video signals in both offline and online tasks. They also showed how their easy-to-use graphical user interface will allow unskilled users to easily access to data recording and classification modules. From this research we realized that it is obvious that greatest recognition software are based upon machine learning approaches and algorithms and therefore these software require a substantial amount of data for dataset training.

Another study in automatic emotion recognition by the speech signal [30] showed how to recognize emotions specifically for spoken utterances. The researchers categorized these utterances levels into two areas: semantic and signal. Semantic level comprises the spoken phrases, which are able to transfer clear reference. Signal level encompasses features, such as pitch, energy, and spectral distortions. They reported the overall accuracy 86.5% for semantic fusion of semantic and signal characteristics of their dataset for seven emotions, such as irritation, joy, anger, fear, disgust, sad, and neutral. They also mentioned that semantic features will not detect all emotions acceptably; however they will support fusion recognition effectively.

From these studies it becomes clear that there are several unsolved issues and a real breakthrough in online and automatic voice emotion recognition systems that have not yet been achieved. We plan to deal with more emotions and increase accuracy in our real-time voice emotion recognition system. In this paper we propose 1) an unobtrusive approach with 2) an objective method that can be verified by researchers, 3) which requires inexpensive and ubiquitous equipment (microphone), and 4) which offers interactive software. Our software uses a microphone and provides the real time conversion of detected vocal intonations into emotional states. Our approach can be applied in any online game-based learning setting that requires additional ways of gathering affective data from the user during learning.

To characterize the novelty of our work, we propose the multimodal framework, which in real-time transforms data from behavioral observations into emotional states. Furthermore, this is applied in educational settings, more precise for training purposes. To our knowledge, these approaches have not yet been integrated in any other frameworks. In this paper, section 2 introduces the FILTWAM framework and its voice emotion recognition part. The method for the proof of concept study of the developed software is described in section 3. Results and discussion are presented in section 4. Section 5 explains the conclusion of this proof of concept study.

2 The FILTWAM Framework

The FILTWAM framework encompasses five functional layers and a number of sub-components within the layers. The five layers are introduced as the: 1) Learner, 2) Device, 3) Network, 4) Application, and 5) Data. Figure 1 illustrates the framework.

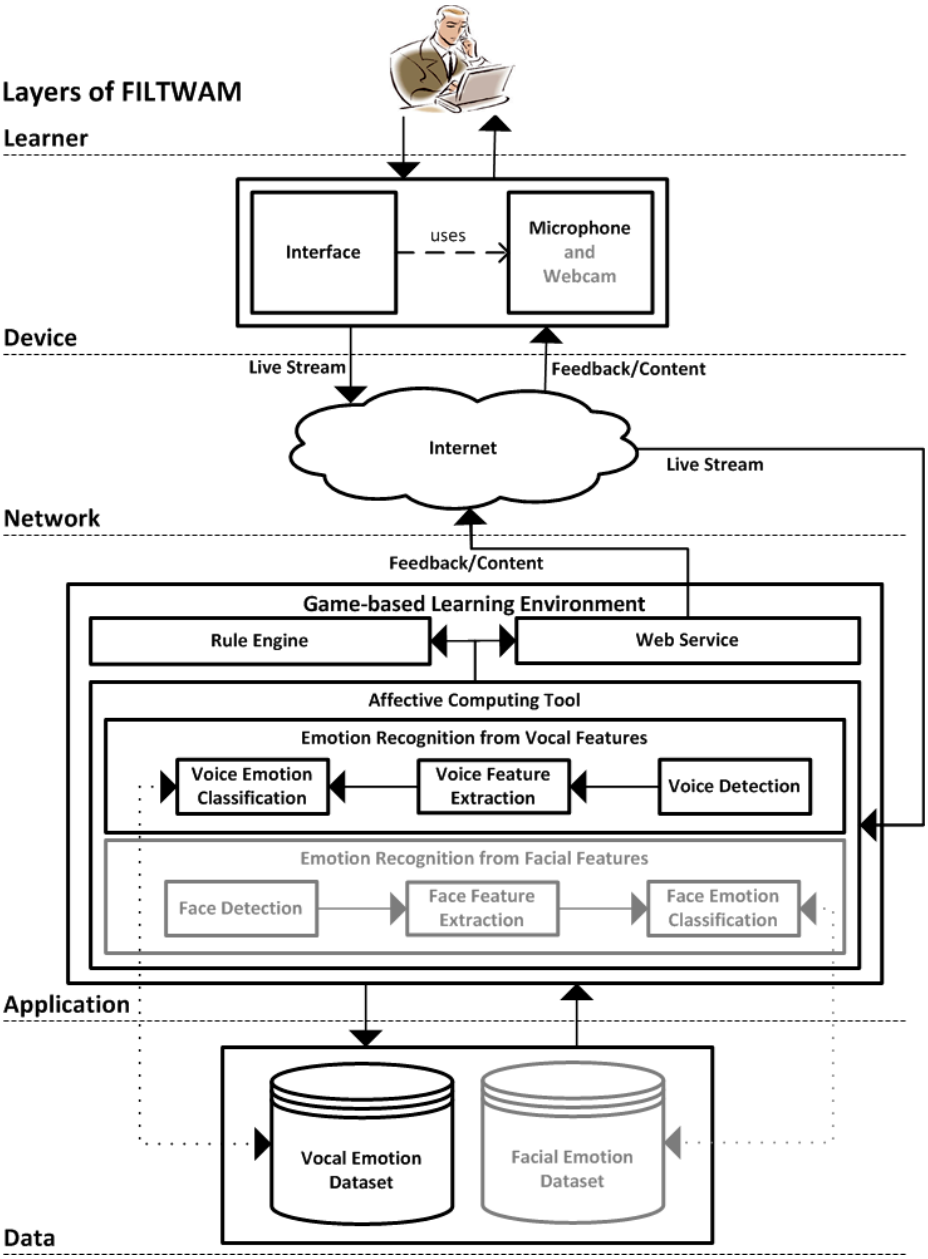

Fig. 1. The FILTWAM framework for voice emotion recognition in an online game-based environment (the face emotion recognition sub-components have been reported in our previous study [31] and are grayed consequently).

2.1 Learner Layer

The learner refers to a subject who uses web-based learning materials for personal development or preparing for an exam.

2.2 Device Layer

The device reflects the learner's workstation, whether part of a personal computer, a laptop, or a smart device, and it includes a webcam and microphone for collecting user data.

2.3 Network Layer

The network uses Internet to broadcast a live stream of the learner and to receive the feedback of the learner.

2.4 Application Layer

The application layer is the most important part of FILTWAM. It consists of game-based learning environment and several sub-components. The game-based learning environment uses a microphone and the vocal emotion recognition technology to facilitate the learning process for the learner. It contains three sub-components named: the affective computing tool, the rule engine, and the web server.

2.4.1 Application Layer

The affective computing tool is the heart of FILTWAM. It processes the facial behavior and vocal intonations data of the learner. It consists of a component for emotion recognition from vocal intonations and facial features. In this paper we confine ourselves to the voice emotion recognition based on the microphone voice streams.

Emotion Recognition from Vocal Features

This component extracts vocal intonations from voices and classifies emotions. It includes three sub-components that lead to the recognition and categorization of a specific emotion.

Voice Detection

The process of emotion recognition from vocal intonations starts at the voice detection component. But we do not necessarily want to recognize the particular

voice; instead we intend to detect a voice and to recognize its vocal emotions. This component divides the received voice signal into meaningful parts that will be used in voice feature extraction and voice emotion classification components.

Voice Feature Extraction

Once the voice is detected, the voice feature extraction component extracts a sufficient set of features from voice of the learner. These features are considered as the significant features of the learner's voice and can be automatically extracted.

Voice Emotion Classification

We adhere to a well-known emotion classification approach that has been introduced by Ekman and has often been used over the past thirty years which focuses on classifying the six basic emotions: sadness, anger, disgust, fear, happiness, surprise (Ekman & Friesen, 1978). However this approach introduced for facial coding systems, but our voice emotion classification component supports classification of these six basic emotions plus the neutral emotion for vocal intonations. This component analyses voice stream and can extract a millisecond feature of each voice stream for its analysis. Currently, we use the naïve Bayes (NB) classifier classification algorithm in FILTWAM. The NB classifier is very fast and appropriate for real-time emotion recognition. Our voice emotion recognition software supports speaker independent recognition approach, which is a general recognition system and therefore its accuracy is lower than the speaker dependent recognition approach that has been reported in [28].

2.4.2 Rule Engine

The rule engine component manages didactical rules and triggers the relevant rules for providing feedback as well as tuned training content to the learner via the device. The game-based learning environment component complies with a specific rule-based didactical approach for the training of the learners. In the future we may possibly use the rule engine of EMERGO, which is a game-based toolkit for delivery of multimedia cases.

2.4.3 Web Service

The web service component transmits the feedback and training content to the learner. At this stage, the learner can receive a feedback based on his/her vocal emotions.

2.5 Data Layer

The data layer is the physical storage of the emotions. It encompasses the vocal emotion dataset, which reflects the intelligent capital of the system. Its records provide a statistical reference for emotion detection.

3 Method and Proof of Concept

Our hypothesis is that data gathered via microphone can be reliably used to unobtrusively infer learners' emotional states. Such emotional states' measurements would allow for the provision of useful feedback during online game-based training of communication skills or any other adaptive or personalized interventions that would enhance the quality and efficacy of e-learning.

3.1 Participants

An e-mail was sent out to employees from the Centre for Learning Sciences and Technologies (CELSTEC) at the Open University Netherlands to recruit the participants for this proof of concept study. The e-mail mentioned the estimated time investment for enrolling in the proof of concept study. Six participants, all employees from CELSTEC (3 male, 3 female; age M=43, SD=9), volunteered to participate in this proof of concept study. By signing an agreement form, the participants allowed us to capture their facial expressions and voice intonations, and to use their data for the proof of concept study. The participants were invited to test the voice emotion recognition module of the affective computing software and watch their face emotions through the face emotion recognition module of the affective computing software; no specific background knowledge was requested. They were told that participation within the proof of concept study might help them to become more aware of their emotions while they were communicating through a microphone and a webcam in the affective computing software.

3.2 Design

Four consecutive tasks were given to the participants. Participants were asked to expose seven basic voice expressions (happy, sad, surprise, fear, disgust, angry, and neutral). Totally, eighty-two voice expressions were requested for all four tasks together. The participants were requested to mimic all the eighty-two emotions once. At the moment, we offer very limited learner support (just a straight forward simple feedback (name of the recognized emotion and the prediction accuracy amount)) to inform the learner whether our current prototype of the affective computing software detects the same 'emotion' as the participant was asked to 'mimic'. For the validation of the software, it is important to know whether its detection is correct. For the learners it is important that they can trust that the feedback is correct.

In the first task participants were asked to speak aloud and use the voice emotion that was shown on the face of the person that was on the image to them. There were 14 images presented subsequently through PowerPoint slides; the participant paced the slides. Each image illustrated a single emotion. All seven basic face expressions were two times present with the following order: happy, sad, surprise, fear, disgust, angry, neutral, happy, sad, et cetera. In the second task, participants were requested to speak aloud the seven basic expressions twice: first, through slides that each presented the keyword of the requested emotion and second, through slides that each presented the keyword and the picture of the requested voice emotion with the following order: angry, disgust, fear, happy, neutral, sad, surprise. For the first and second tasks, they could improvise and use their own texts. The third task presented 17 slides with the text transcript (both sender and receiver) taken from a good-news conversation. The text transcript also included instructions which voice expression should accompany the current text-slide. Here, participants were requested to read and speak aloud the sender text of the 'slides' from the transcript and deliver the accompanying voice expression. The forth task with 37 slides was similar to task 3, but in this case the text transcript was taken from a bad-news conversation. The transcripts and instructions for tasks 3 and 4 were taken from an existing OUNL training course [32] and a communication book [33].

3.3 Test Environment

All tasks were performed on a single Mac machine. The Mac screen was separated in three panels, top-left, top-right, and bottom. The participants could watch their facial expressions in the face emotion recognition module of the affective computing software at the top-left panel, they could watch their analyzed voice expressions in the voice emotion recognition module of the affective computing software at the top-right panel, while they were performing the tasks using a PowerPoint file in the bottom panel. An integrated webcam with a microphone and a 1080HD external camera were used to capture and record the emotions of the participants as well as their actions on the computer screen. The affective computing software including the face emotion recognition module and the voice emotion recognition module used the webcam with the microphone to capture and recognize the participants' emotions, while Silverback usability testing software version 2.0 used the external camera to capture and record the complete experimental session. Figure 2 demonstrates an output of both modules of the software and an experimental session for Task 4.

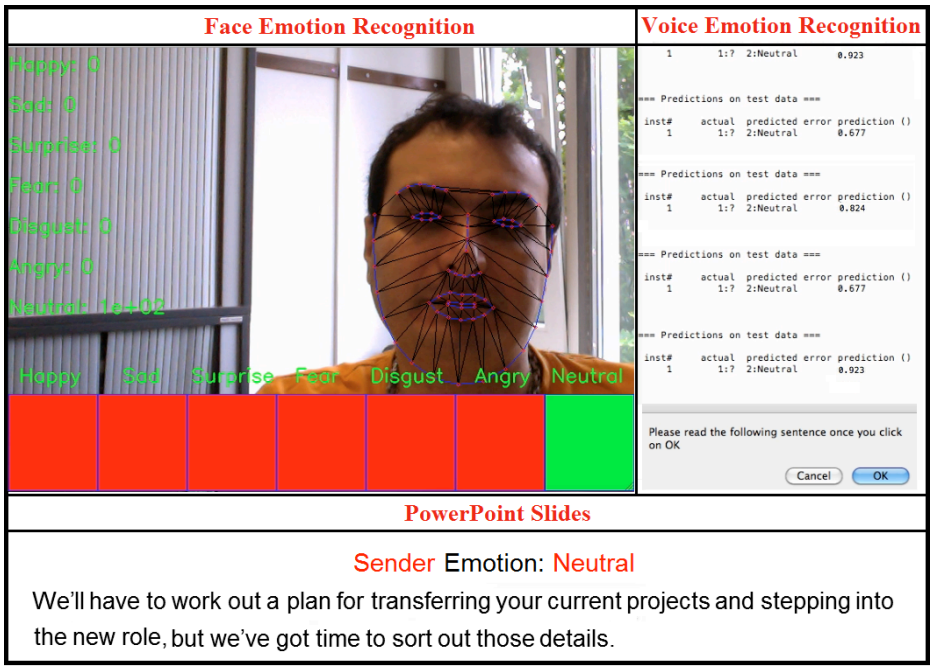

Fig. 2. The main researcher in task 4, the affective computing software including the face emotion recognition module (top-left) and the voice emotion recognition module (top-right), and the PowerPoint slide (bottom) during the experimental session.

3.4 Measurement Instruments

A self-developed online questionnaire was used to collect participants' opinions when carrying out the requested tasks. All opinions were collected using items on a 7- point scale format with possible scores: 1) completely disagree, 2) disagree, 3) mildly disagree, 4) neither disagree nor agree, 5) mildly agree, 6) agree, and 7) completely agree. Participants' opinions were gathered for: 1) perceived difficulty to mimic the requested emotions in the given tasks, 2) perceived feedback of the given feedback to mimic the emotions in the given tasks, 3) perceived instructiveness of the instructions for the given tasks, 4) perceived attractiveness of the given tasks, and 5) perceived concentration on the given tasks. Participants were also asked to report their self-assurance on 1) being able to mimic the requested emotions in the given tasks and 2) their acting skills on a similar 7-point Likert scale.

3.5 Procedure

Each participant signed the agreement form before his or her session of the proof of concept study was started. They individually performed all four tasks in a single session of about 45 minutes. The session was conducted in a completely silent room with a good lighting condition. The moderator of the session was present in the room, but did not intervene. All sessions were conducted in three consecutive days. The participants were requested not to talk to each other in between sessions so that they could not influence each other. The moderator gave a short instruction at the beginning of each task. For example, participants were asked to show mild and not too intense expressions while mimicking the emotions. All tasks were recorded and captured by both the face emotion recognition module and the voice emotion recognition module of the affective computing software. After the session, each participant filled out an online Google questionnaire form to help us to collect feedback for improvement the tasks and the software and the setting of the study.

4 Results and Discussion

4.1 The Dataset

The vocal dataset was created in the Centre for Learning Sciences and Technologies (CELSTEC) at the Open University Netherlands. Currently FILTWAM propose a dataset in English language with 700 occurrences (100 for each category). It supports seven basic emotions, such as happy, sad, surprise, fear, disgust, angry, and neutral. The dataset was prepared by 10 non-expert speakers.

4.2 The Voice Feature Extraction and Classification

The voice emotion recognition software of the FILTWAM requires reaching to a better accuracy level when all the seven basic emotions integrate in a single dataset. The voice feature extraction and the voice emotion classification could be improved to reach to higher accuracy when an equal number of actors and actresses are recruited to train the new dataset. A combination of the voice emotion recognition module of the affective computing tool of the FILTWAM framework with the face emotion recognition module of it may help us to improve the possible recognition rates.

4.3 Validation Results of the Software

In this paper we report the validation of the voice emotion recognition module of our affective computing software, all tasks, and six basic emotions as well as neutral emotion. For the first validation results, we only selected two emotional categorize

(happy and neutral) from our dataset. Table 1 shows the results of the requested emotions from participants and compares the results with recognized emotions by the voice emotion recognition module of the affective computing software.

Table 1. Validation results of the software for only happy and neutral emotions for task 1, task 2, task3, and task 4 simultaneously.

Requested Emotion	Recognized Emotion by the Voice Emotion Recognition Software							
	Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
Happy	26	----	----	----	----	----	10	36
	72.2%	----	----	----	----	----	27.8%	100%
Neutral	74	----	----	----	----	----	250	324
	23%	----	----	----	----	----	77%	100%

Each requested emotion is separated in two rows that intersect with the recognized emotions by the voice emotion recognition software. The first row indicates the number of occurrences of the recognized emotion and the second row displays the percentage of each recognized emotion. The red numbers are the correctly classified emotions and therefore the accuracy of the voice emotion recognition module, while the black numbers are the incorrectly classified emotions. The best recognized emotion in this case is still neutral 77% followed by happy 72.2%. Therefore the overall accuracy is 74.6%. In accordance with [34], we realized that the classification accuracy decreases with the number of emotional categories in our current dataset. Therefore when we inserted sad, surprise, fear, disgust, and angry to our dataset, the classification accuracy decreased (see Table 2).

Table 2. Validation results of the software for all the seven emotion for task 1, task 2, task3, and task 4 simultaneously.

Requested Emotion	Recognized Emotion by the Voice Emotion Recognition Software							
	Happy	Sad	Surprise	Fear	Disgust	Angry	Neutral	Total
Happy	16	3	3	1	7	0	18	48
	33.3%	6.3%	6.3%	2.1%	14.5%	0.0%	37.5%	100%
Sad	3	3	0	0	1	0	17	24
	12.5%	12.5%	0.0%	0.0%	4.2%	0.0%	70.8%	100%
Surprise	3	2	2	2	4	2	9	24
	12.5%	8.3%	8.3%	8.3%	16.8%	8.3%	37.5%	100%
Fear	2	3	1	1	1	1	15	24
	8.2%	12.5%	4.2%	4.2%	4.2%	4.2%	62.5%	100%
Disgust	2	4	0	1	2	2	13	24
	8.3%	16.7%	0.0%	4.2%	8.3%	8.3%	54.2%	100%
Angry	10	3	1	0	3	2	5	24
	41.7%	12.5%	4.2%	0.0%	12.5%	8.3%	20.8%	100%
Neutral	21	28	0	6	0	8	261	324
	6.5%	8.5%	0.0%	1.9%	0.0%	2.5%	80.6%	100%

The achieved overall accuracy of the software between the requested emotions and the recognized emotions assuming uniform distribution of emotions is the average of

the diagonal: 22.2% (based on Table 2). The best recognized emotion is still neutral 80.6% followed by happy 33.3%, sad 12.5%, surprise 8.3%, disgust 8.3%, angry 8.3%, and fear 4.2%. Currently there are three reasons for the obtained false results: 1) the malfunctioning of the software, 2) the participants were unable to mimic the requested emotions accurately, and 3) the accuracy of the trained voice emotion dataset is less than 50% now. We know that we are in a good track, but we are not sure yet what will be changed if we recruit several actors and actresses who are capable of mimicking the requested emotions accurately to train our dataset. There will be more investigation on these issues in our feature research and development.

4.4 Summary of the Measurement Parameters

In this section we report the results of the online questionnaire designed to measure the seven parameters (difficulty, feedback, self-assurance, instructiveness, attractiveness, concentration, and acting skills) mentioned in the section 3.4. The participants completed the questionnaire regarding the voice emotion recognition experimental session. Table 3 summarizes the measurement parameters that filled out by the participants through the online questionnaire investigation.

Table 3. Summary of the measurement parameters.

Questions	Answers by the Participants								
		1	2	3	4	5	6	7	Total
	Difficulty It was easy for me to mimic the requested emotions in the given tasks	----	17%	17%	9%	22%	35%	----	
	Feedback The feedback did help me to mimic the emotions in the given tasks	----	8%	4%	25%	25%	38%	----	
	Self-assurance I am confident that I was able to mimic the requested emotions in the given tasks	3%	3%	58%	10%	10%	16%	----	
	Instructiveness The instructions for the given tasks were clear to me	----	----	----	----	33%	59%	8%	100%
	Attractiveness The given tasks were interesting	----	----	----	----	73%	27%	----	
	Concentration I could easily focus on the given tasks and was not distracted by other factors	----	----	----	4%	12%	50%	34%	
	Acting skills I regard myself as a good actor	----	33%	17%	33%	----	17%	----	

1= Completely disagree, 2= Disagree, 3= Mildly disagree, 4= Neither disagree nor agree, 5= Mildly agree, 6= Agree, and 7= Completely agree

The observation of Table 3 shows that self-assurance parameter is less for tasks 1 and 2 as compared to the other tasks. It is easy to realize that the participants don't regard themselves as actors. The difficulty parameter contrary to our expectations doesn't show any differences and all tasks seem moderately difficult. Clear instructions and feedback parameters are moderately helpful. All the tasks were interesting for the participants to do. Finally concentration parameter indicates no distraction during performance.

5 Conclusion

We have examined a proof of concept of software for real-time voice emotion recognition that is part of our FILTWAM framework. The overall accuracy of our voice emotion recognition software based on the requested emotions and the recognized emotions for two emotions (happy and neutral) is 74.6 %. However we have lower values for an extended set of the emotions. This is in accordance with [28, 34] in which we expect low accuracy for completely normal low-intensity emotions in online recognition and decreasing the classification accuracy with the number of emotional categories. This issue requires further investigation and improvement. Participants indicated that tasks 1 and task 2 were more difficult for them than tasks 3 and 4. Tasks 1 and 2 included a bigger variety of requested emotions. In combination with questionnaire results indicating participants' low self-confidence on being a good actor, this might hint at asking professional actors for a next validation study of the software. Furthermore, the findings with the questionnaire indicate that there is some room left for improving the tasks offered to participants.
We will further investigate possibly software malfunctioning and might include more training sessions for participants to improve their skills for being able to mimic the requested emotions more accurately. The FILTWAM framework aims at real-time interpretation of emotional behavior into emotional states and can be applied in game-based learning in general and seems especially useful for communication skills game training in particular. We aim to use the validated framework for communication skills game training purposes in our further research and development. We will further integrate the face emotion recognition module and the voice emotion recognition module. The current FILTWAM framework, including the face emotion recognition software and the voice emotion recognition software, is an advanced human-computer interaction setting that can be integrated with existing game-based learning environments. For our research, FILTWAM will be integrated with the EMERGO-platform [10].

Acknowledgments. We thank our colleagues who participated in the voice emotion recognition proof of concept study. This research is sponsored by The Netherlands Laboratory for Lifelong Learning (NELLL) of the Open University of the Netherlands.

References

1. Anaraki, F.: Developing an Effective and Efficient eLearning Platform. International Journal of The Computer, the Internet and Management. Vol. 12(2):57-63. (2004)
2. Nagarajan, P., Wiselin, G. J.: ONLINE EDUCATIONAL SYSTEM (e- learning). International Journal of u- and e- Service, Science and Technology. Vol. 3(4):37-48. (2010)
3. Norman, G.: Effectiveness, efficiency, and e-learning. Journal of Advances in Health Sciences Education. Vol. 13 (3):249-251. (2008)
4. Ebner, M.: E-Learning 2.0 = e-Learning 1.0 + Web 2.0?. The Second International Conference on Availability, Reliability and Security (ARES). p. 1235-1239. (2007)
5. Hrastinski, S.: Asynchronous & synchronous e-learning. Educause Quarterly. Vol. 31(4):51-55. (2008)
6. Kelle, S., Sigurðarson, S., Westera, W., Specht, M.: Game-Based Life-Long Learning. In G. D. Magoulas (Ed.), E-Infrastructures and Technologies for Lifelong Learning: Next Generation Environments. Hershey, PA: IGI Global. p. 337-349. (2011)
7. Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., Boyle, J. M.: A systematic literature review of empirical evidence on computer games and serious games. Computers and Education. September. Vol. 59(2):661-686. (2012)
8. Reeves, B., Read, J.L.: Total engagement: Using games and virtual worlds to change the way people work and business compete. Boston. Harvard Business Press. (2009)
9. Gee, J.P.: What video games have to teach us about learning and literacy. New York: Palgrave Macmillan. (2003)
10. Nadolski, R. J., Hummel, H. G. K., Van den Brink, H. J., Hoefakker, R., Slootmaker, A., Kurvers, H., Storm, J.: EMERGO: methodology and toolkit for efficient development of serious games in higher education. Simulations & Gaming. Vol. 39(3):338-352. (2008)
11. Bahreini, K., Nadolski, R., Qi, W., Westera, W.: FILTWAM - A Framework for Online Game-based Communication Skills Training - Using Webcams and Microphones for Enhancing Learner Support. In P. Felicia (Ed.), The 6th European Conference on Games Based Learning (ECGBL). Cork, Ireland. p. 39-48. (2012)
12. Avidan, S., Butman, M.: Blind vision. European Conference on Computer Vision. Vol. 3953:1–13. (2006)
13. Bashyal, S., Venayagamoorthy, G.K.: Recognition of facial expressions using Gabor wavelets and learning vector quantization. Engineering Applications of Artificial Intelligence. (2008)
14. Chibelushi, C. C., Bourel, F.: Facial expression recognition: a brief tutorial overview. Available Online in Compendium of Computer Vision. (2003)
15. Ekman, P., Friesen, W. V.: Facial Action Coding System: Investigator's Guide. Consulting Psychologists Press. (1978)
16. Kanade, T.: Picture processing system by computer complex and recognition of human faces. PhD thesis. Kyoto University, Japan. (1973)
17. Li, S. Z., Jain, A. K.: Handbook of Face Recognition Second Edition. ISBN 978-0-85729-931-4. Springer-Verlag, London. (2011)
18. Petta, P., Pelachaud, C., Cowie, R.: Emotion-Oriented Systems. The Humaine Handbook. Springer-Verlag. Berlin. (2011)
19. Chen, L.S.: Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction. University of Illinois at Urbana-Champaign. PhD thesis. (2000)
20. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems. Vol. 42(3-4):143-166. (2003)

21. Sebe, N., Cohen, I. I., Gevers, T., Huang, T. S.: Emotion recognition based on joint visual and audio cues. International Conference on Pattern Recognition. Hong Kong. p. 1136-1139. (2006)
22. Song, M., Bu, J., Chen, C., Li, N.: Audio-visual based emotion recognition: A new approach. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2. (2004)
23. Subramanian, R., Staiano, J., Kalimeri, K., Sebe, N., Pianesi, F.: Putting the Pieces Together: Multimodal Analysis of Social Attention in Meetings. ACM Multimedia. Firenze. Italy. (2010)
24. Zeng, Z., Pantic, M., Roisman, G. I., Huang, T. S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31(1):39–58. (2009)
25. Sebe, N.: Multimodal Interfaces: Challenges and Perspectives. Journal of Ambient Intelligence and Smart Environments. January. Vol. 1(1):23-30. (2009)
26. Pekrun, R.: The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. Journal of Applied Psychology. Vol. 41:359–376. (1992)
27. Hager, P. J., Hager, P., Halliday, J.: Recovering Informal Learning: Wisdom, Judgment And Community. Springer. (2006)
28. Vogt, T. André, E. Bee, N.: EmoVoice - A framework for online recognition of emotions from voice. In Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems. (2008)
29. Wagner, J. Lingenfelser, F. Andre, E.: The Social Signal Interpretation Framework (SSI) for Real Time Signal Processing and Recognitions. In Proceedings of INTERSPEECH. Florence, Italy. (2011)
30. Schuller, B., Manfred, L., Gerhard, R.: Automatic emotion recognition by the speech signal. Institute for Human-Machine-Communication. Technical University of Munich. 80290. (2002)
31. Bahreini, K., Nadolski, R., Westera, W.: FILTWAM - A Framework for Online Affective Computing in Serious Games. The 4th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES'12). Procedia Computer Science. Genoa, Italy. Vol. 15:45-52. (2012)
32. Lang, G., van der Molen, H. T.: Psychologische gespreksvoering. Open University of the Netherlands. Heerlen, The Netherlands. (2008)
33. Van der Molen, H. T., Gramsbergen-Hoogland, Y. H.: Communication in Organizations: Basic Skills and Conversation Models. ISBN 978-1-84169-556-3. Psychology Press, New York. (2005)
34. Dai, K., Harriet J. F., MacAuslan, J.: Recognizing emotion in speech using neural networks. Telehealth and Assistive Technologies. p.31-38. (2008)