Educ Inf Technol (2016) 21:1367–1386
DOI 10.1007/s10639-015-9388-2

Towards real-time speech emotion recognition for affective e-learning

Kiavash Bahreini[1] · **Rob Nadolski**[1] · **Wim Westera**[1]

Published online: 15 April 2015

Abstract This paper presents the voice emotion recognition part of the FILTWAM framework for real-time emotion recognition in affective e-learning settings. FILT WAM (Framework for Improving Learning Through Webcams And Microphones) intends to offer timely and appropriate online feedback based upon learner's vocal intonations and facial expressions in order to foster their learning. Whereas the facial emotion recognition part has been successfully tested in a previous study, the here presented study describes the development and testing of FILTWAM's vocal emotion recognition software artefact. The main goal of this study was to show the valid use of computer microphone data for real-time and adequate interpretation of vocal intonations into extracted emotional states. The software that was developed was tested in a study with 12 participants. All participants individually received the same computer-based tasks in which they were requested 80 times to mimic specific vocal expressions (960 occurrences in total). Each individual session was recorded on video. For the validation of the voice emotion recognition software artefact, two experts annotated and rated participants' recorded behaviours. Expert findings were then compared with the software recognition results and showed an overall accuracy of Kappa of 0.743. The overall accuracy of the voice emotion recognition software artefact is 67 % based on the requested emotions and the recognized emotions. Our FILTWAM-software allows to continually and unobtrusively observing learners' behaviours and transforms these behaviours into emotional states. This paves the way for unobtrusive and real-time capturing of learners' emotional states for enhancing adaptive e-learning approaches.

Keywords Speech interaction · Affective computing · Speech emotion recognition · Real-time software development · Evaluation methodology · Empirical study of user behaviour · E-learning · Microphone

✉ Kiavash Bahreini
kiavash.bahreini@ou.nl

[1] Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences, Open University of the Netherlands, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands

1 Introduction

Affective computing is an emerging research domain that focuses on natural interactions between humans and computers. There is a need for applications that can recognize human emotions to facilitate smoother interaction between humans and computers. Recognizing emotions in daily speech on a real-time basis is a difficult task for computers and constitutes a challenging area of research and software development (López-Cózar et al. 2011). Different areas of e-learning can benefit from affective user data since emotional states are relevant for learning processes (Bachiller et al. 2010). It is widely accepted that emotions are significant factors in any learning process, because they influence information processing, memory and performance (Pekrun 1992). Delivering feedback to learners becomes more personalized when emotional states are taken into account. Similarly, feedback based on emotional states may be beneficial and can enhance learners' awareness of their own behaviour. It is important for learners to learn how to express the correct emotion at the right time. Appropriate feedback can guide the alignment between emotions and the message content during communication skills training. Nowadays, learners regularly use web-based applications for communicating, working and learning with peers in distributed settings (Ebner 2007). In the past, detecting learner emotions has not been well developed in such settings. More recently, however, various studies with different accuracy levels have become available (Happy et al. 2013; Liu et al. 2011).

In this study, we introduce our developed voice emotion recognition software artefact of our FILTWAM framework that can be used in any e-learning settings. We additionally describe its practical applications and present results of its first evaluation. The theoretical and conceptual aspects of the FILTWAM framework have been described in our previous study (Bahreini et al. 2012). Although our framework allows for both facial and vocal emotion detections and can generate timely feedback based upon learner's facial and vocal expressions, we restrict ourselves here to voice emotion recognition and provide empirical data for this goal. Our software artefact is able to recognize the following emotions: happiness, surprise, anger, disgust, sadness, fear, and neutral. For purposes of this study, we focused on communication skills training. We used webcams and microphones to offer an easy and readily available means of collecting data for emotion recognition. Computer microphones allow for more natural interactions with the e-learning applications. They can be used for the nonintrusive and continuous collection of emotional user data (e.g., from speech). In this study, we used a common computer microphone for gathering uninterrupted affective user data in an e-learning context.

This paper presents 1) an unobtrusive approach 2) with an objective method that can be verified by researchers 3) involving inexpensive and ubiquitous equipment (microphone), and 4) which offers interactive software with user-friendly interface. Moreover, this paper describes the theoretical contribution of the study for voice emotion recognition in e-learning environments. Particularly for emotion recognition software that is real-time, unobtrusive, and functional in a continuous learning context. It also describes our software artefact in terms of its innovative characteristic (real-time) and its high level of reliable detection. Section 2 focuses on a review of relevant literature. Section 3 describes the FILTWAM framework and its voice emotion recognition software artefact. The methodology for evaluating the software artefact study is described in

section 4. Results from the empirical study of user behaviour are explained in section 5. Ethical implications are stated in section 6. Discussion and findings are described in section 7. Summary, outlook, and suggestions for future work are discussed in section 8. Section 9 provides conclusions related to the research.

2 Literature review

Existing methods for collecting affective learner data, such as physiological sensors or questionnaires, are more limited, and they inevitably disrupt the process of learning (Feidakis et al. 2011). Most of the research deals with using emotions for adapting learning content or learning tasks. This insight has led to the research and development of affective tutoring systems (Sarrafzadeh et al. 2008). We expand the application context of our voice emotion recognition software artefact to communication skills training in e-learning settings using a microphone.

Previous research has shown that the automatic software development for online emotion recognition from speech fragments is developed for general purposes and not for e-learning environments that require specific settings and user modelling. For example, Wagner and colleagues (Wagner et al. 2011) implemented a social signal interpretation framework for real-time signal processing and recognition. Wagner and colleagues (Wagner et al. 2013) developed a tool for an automatic detection and interpretation of social signals of speech. Jones and Sutherland (2008) developed a system for human-robot interaction for feedback provision. Vogt and colleagues developed an approach (Vogt et al. 2008a, b) for automatic analysis of speech fragments enabling unobtrusive gathering of affective learner data in online e-learning settings. Their approach makes it possible to extract vocal intonations and map them onto emotional states, and is the approach we followed in our study.

An important factor for successful teaching is the teacher's ability to recognize and respond to the affective states of students. In e-learning, just as with conventional classroom learning, both the cognitive dimension and its connection with emotion are important. Emotion software system developed for e-learning could significantly increase learner performance by adapting the software to the emotional state of the learner (Sarrafzadeh et al. 2008). In e-learning, the limited availability of teachers caused many studies on affective computing. Affective computing helps overcome the tendency to disregard emotions and also incorporates human-like capabilities of interpretation and generation of affect features (Jianhua et al. 2005). Beale and Creed (2009) investigated the impact of embodied agents' emotions on users' behaviour. They determined that the co-learning agent with an emotion-enabled characteristic would enhance interactions in the learning and education domain. Our developed software design is based upon several earlier studies (Chibelushi and Bourel 2003; Ekman and Friesen 1978; Vogt et al. 2008a, b; Wagner et al. 2013). We investigated whether or not data gathered via microphone (real-time voice recognition) can be used to reliably and unobtrusively gather learners' emotional states. Such emotional state measurements are beneficial for providing useful learning feedback during online communication skills training.

Several studies in the past (Bozkurt et al. 2009; Neiberg et al. 2006; Schröder 2009) have shown that automatic emotion recognition from speech takes speech fragments as input data. It is generally acknowledged that recognizing which features are indicative of emotional

states and capturing them as speech fragments is a complex task (Pfister, and Robinson 2011). Each emotion occurring in each human-computer interaction is spontaneous, making it difficult to distinguish the acoustic features within each interaction (Vogt et al. 2008a, b). The task is further complicated by the fact that there is not an unambiguous answer indicating how a given speech sample should be classified with a specific emotion. The speech sample could easily include more than one emotion, making it difficult to distinguish separate emotions within one fragment. Furthermore, emotions expressed in natural speech are more challenging to identify compared to acted speech (Batliner et al. 2003). We follow the preceding approaches that explained the challenges of capturing emotions in speech fragments and integrate speech fragments as input data into our developed software artefact.

Feidakis and colleagues (Feidakis et al. 2011) explain how to measure emotions for intelligent tutoring systems (ITS). They categorized emotion measurement tools into three areas: psychological, physiological, and motor-behavioural. Psychological tools are self-reporting tools for capturing the subjective experience of emotions of users. Physiological tools are sensors that capture an individual's physiological responses. Motor-behaviour tools use special software to measure behavioural movements captured by the PC cameras, mouse or keyboard. Many practical applications would considerably increase performance if they could adjust to the emotional state of the user. When equipped with affective computing software artefact, an ITS can be turned into an affective tutoring system (ATS). Hence, a computer application (e.g., our developed software artefact) that is able to recognize users' vocal emotions can improve feedback to learners without involvement of a human teacher. There is a growing body of research on ATS stressing the importance of an approach using vocal expressions for deriving emotions (Ben Ammar et al. 2010; Sarrafzadeh et al. 2008), and this approach has been incorporated into our own software design.

3 The FILTWAM framework

The FILTWAM framework aims to improve learners' communication skills training by providing timely and adequate feedback to them exploiting their state data. The data are gathered through webcam and microphone when interacting with online training materials in an e-learning environment. This framework consists of five layers and a number of sub-components within the layers. The five layers are presented as the: 1) Learner, 2) Device, 3) Data, 4) Network, and 5) Application. Figure 1 illustrates the framework.

3.1 Learner layer

The learner refers to a subject who uses web-based learning materials for personal development, preparing for an exam, or aims at interacting with an affective-enabled e-learning environment. The learner is a lifelong learner who is positively biased towards the paradigm of informal learning and who prefers to study at his own pace, place, and time.

3.2 Device layer

From technical perspective the device layer is the most important part of FILTWAM. This layer indicates the learner's hardware configuration, whether part of a personal

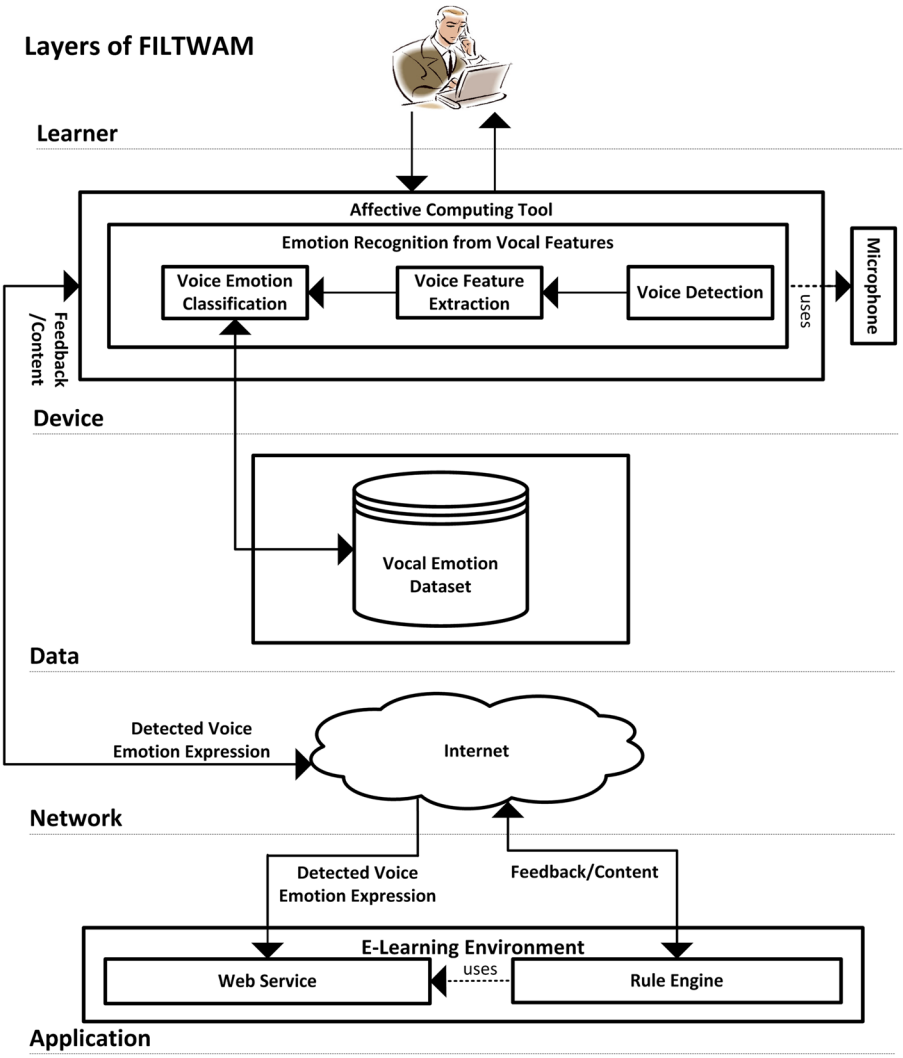

Fig. 1 FILTWAM framework for speech interaction and real-time voice emotion recognition in affective e-learning environments

computer, a laptop, or a smart device. It is supposed to include a microphone for collecting user voice data. It contains one sub-component that is called: the affective computing tool.

3.2.1 Affective computing tool

The affective computing tool is the heart of FILTWAM. It processes the vocal intonations data of the learner. It consists of a component for emotion recognition from vocal features. The emotion recognition from vocal features uses the voice streams from the microphone. The user interface of this tool sends the detected voice expression of the

learner to the web service sub-component of the e-learning environment component in the application layer through the network layer. It receives the feedback and the content that are provided by the rule engine component in the application layer through the Internet component in the network layer. We developed and tested our software artefact using two existing tools: 1) the Praat[1] tool, which is a tool for speech analysis and 2) the openSMILE[2] tool, which is an open source tool for audio feature extraction.

3.3 Emotion recognition from vocal features

Based on pauses of 1 s, this component divides speech into meaningful segments representing particular emotions for voice emotion recognition. It extracts vocal features from voices and classifies emotions. It includes three sub-components that lead to the recognition and categorization of a specific emotion. Voice detection, voice feature extraction, and voice emotion classification are placed in this component.

3.4 Voice detection

The process of emotion recognition from vocal features starts at the voice detection component. But we do not necessarily want to recognize the particular voice; instead we intend to detect a voice to recognize its vocal emotions. We start with segmenting the input speech into significant fragments to be used in the classification function later. We are not interested in statistical calculation of the words as they are too short for this purpose. Instead we use sentence level as an appropriate fragment for spontaneous speech. This allows us to provide a reliable statistical analysis for feature extraction in each separate fragment.

3.5 Voice feature extraction

Once the voice is detected, the voice feature extraction component extracts a sufficient set of feature points of the learner. These feature points are considered as the significant features of the learner's voice and can be automatically extracted. We find several features of the speech signal, such as pitch and energy to characterize the emotions. We then put the features into vectors. Finally, we consider the sequences of labelled features in the generated vectors and try to find the changes between the vectors.

3.6 Voice emotion classification

We adhere to a well-known emotion classification approach that has frequently been used over the past three decades which focuses on classifying the six basic emotions (Ekman and Friesen 1978). Our vocal emotion classification component supports classification of these six basic emotions and the neutral emotion. This component analyses voice stream and can extract a millisecond feature of each voice stream for its analysis. Currently, we use the sequential minimal optimization (SMO[3]) classifier of

[1] http://www.fon.hum.uva.nl/praat/
[2] http://opensmile.sourceforge.net/
[3] http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html

WEKA[4] software, which is a software tool for data mining. 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 (Vogt et al. 2008a, b).

3.7 Data layer

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

3.8 Network layer

The network uses Internet to broadcast a detected voice emotion expression of the learner's speech fragments. It broadcasts the feedback and the content of the learner provide by the rule engine component in the application layer.

3.9 Application layer

The application layer consists of e-learning environment and two sub-components. This layer is responsible for managing the received data from the learner and for providing the appropriate feedback and content.

3.10 E-learning environment

The e-learning environment component uses the detected voice emotion expression data of the learner to facilitate the learning process. Its sub-components named: the rule engine component and the web service component.

3.11 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 e-learning environment component complies with a specific rule-based didactical approach for the training of the learners. The web service component transfers the feedback and training content to the learner. At this stage, the learner can receive a feedback based on his or her vocal expressions.

3.12 Web service

The web service component receives the learner's data and makes the data available to the rule engine component.

[4] http://www.cs.waikato.ac.nz/ml/weka

4 Evaluation methodology

For testing our hypothesis mentioned in the introduction section, we have setup four evaluation tasks, directed to the mastery of communication skills.

4.1 Participants

Twelve participants, all employees from the Welten Institute (7 male, 5 female; age M=38, SD=9.7) volunteered to participate in the study. By signing an agreement form, the participants allowed us to record their voice intonations, and to use their data anonymously for future research. The participants were invited to test the voice emotion recognition software artefact; no specific background knowledge was requested.

4.2 Design

Four consecutive tasks were given to the participants, all requiring them to loudly expose seven basic voice expressions. Totally, 80 voice expressions were requested for all four tasks together. All materials were in English language. The learning goal in the current study is that participants become more aware of their own emotions when they deliver their messages. In the first task, participants were asked to loudly mimic the emotion that was presented on the image shown to them. There were 14 images, all were taken from the face of the first author of this paper, presented subsequently through PowerPoint slides; the participant paced the slides. Each image illustrated a single emotion. All seven basic voice expressions were two times present with the following order: happy, sad, surprise, fear, disgust, anger, neutral, happy, et cetera. In the second task, participants were requested to mimic the seven voice expressions twice (14 times in total): 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 emotion with the following order: anger, disgust, fear, happy, neutral, sad, surprise. The third task presented 16 slides linked with a narrative of the text transcripts (both sender and receiver) taken from a good-news conversation. The text transcript also included instructions what vocal expression should accompany the current text-slide. Here, participants were requested to read and speak aloud and mimic the sender text of the ‘slides’ from the transcript. The forth task with 36 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 Open University of The Netherlands (OUNL) training course (Lang and van der Molen 2008) and a communication book (Van der Molen and Gramsbergen-Hoogland 2005). All tasks involve the training of their communication skills. The participants were requested to loudly mimic all the emotions once. At the moment we offer very limited learner support including the predicted name of the emotion and the predicted accuracy number. We inform the learner whether our current prototype of the voice emotion recognition software artefact detects the same ‘emotion’ as the participant was asked to ‘mimic’. For the validation of the software artefact, 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.

4.3 Test environment

Participants performed these tasks individually on a single computer. Figure 2 shows a screen shot of a session for one of the tasks. As presented in the figure, the requested emotion on the PowerPoint slide is 'neutral' and the recognized emotion of the whole sentence is predicted to be 'neutral' by about 92 % accuracy. The screen was separated in two sections, left and right. The participants could watch their vocal expressions feedback generated by the voice emotion recognition software artefact of the affective computing tool at the right section, while they were performing the tasks using the PowerPoint slides in the left section. An integrated webcam and a 1080HD external camera were used to capture and record all the sessions. The voice emotion recognition software artefact used the microphone of the computer to capture, analyse, and recognize the participants' emotions, while Silverback usability testing software (screen and audio recording software) version 2.0 used the external camera to record the complete session. Raters for validating our voice emotion recognition software artefact used the converted audio (wav) files from the video (mov) files.

4.4 Gathering participants' opinions

We have developed an online questionnaire to collect participants' opinion. We requested the participants to report their self-assurance through the questionnaire. All opinions were collected using items on a 7- point Likert scale format (1 = completely disagree, 7 = completely agree). Participants' opinions about their tasks were gathered for: 1) difficulty to mimic the requested emotions, 2) quality of the given feedback 3) clarity of the instructions 4) the attractiveness of the tasks, and 5) their concentration on the given tasks. Moreover, we measured participants' self-assurance by their two 7-point Likert scale items 1) being able to mimic the requested emotions and 2) being able to act.

4.5 Procedure

Each participant signed the informed consent before his or her session of the study was started. Participants individually performed all four tasks in a single session of about

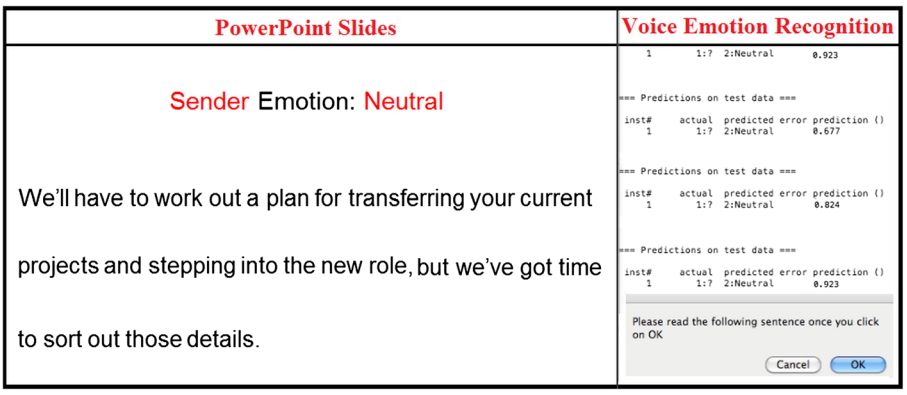

Fig. 2 Task 5 and the voice emotion recognition software artifact during the experimental session

50 min. The session was conducted in a silent room with good lighting conditions. The moderator of the session was present in the room, but did not intervene, except for providing standard instructions. All sessions were conducted in five 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 say mild and not too intense expressions while mimicking the emotions. Directly after the session, each participant filled out the online questionnaire to gather participants' opinions about their learning and the setup of the study.

4.6 Validation

Two raters analysed the recorded wav files and carried out the validation of the software output. First rater is a PhD employee at the Psychology Department of the Open University of the Netherlands and the second rater is a lecturer at the Computer and Electrical Engineering Department of IAU University of Tehran. Both raters individually rated the emotions of the participants' in the recorded voice streams. Both raters are familiar and skilled with voice and speech analysis. Raters overall task was to recognize and rate the recorded voice files for vocal emotion recognition of the participants.

Firstly, they received an instruction package for doing ratings of one of the participants' emotions in 80 different wav files. Secondly, both raters participated in a remotely training session where ratings of the participant were discussed to identify possibly issues with the rating task and to improve common understanding of the rating categories. Thirdly, raters resumed their individual ratings of participants' emotions in the complete voice streams. Fourthly, they participated in a negotiation session where all ratings were discussed to check whether negotiation about dissimilar ratings could lead to similar ratings or to sustained disagreement. Finally, the final ratings resulting from this negotiation session were taken as input for the data analysis.

The data of the training session were also included in the final analysis. The raters received: 1) a user manual, 2) 960 wav files of all 12 participants, 3) an instruction guide on how to recognize the emotions from the audio files, and 4) an excel file with 960 records; each of which represented the name of the audio file and requested for the possible emotion. After the raters rated the dataset, the main researcher of this study compared their results with the software recognition results using IBM SPSS[5], which is a predictive analytics software and R[6], which is a free software environment for statistical computing.

5 Empirical study of user behaviour and results

In this section we report the outcomes of the study. We will first present the comparison of the requested emotions and the recognized emotions by the voice emotion recognition software artefact. Next we will present the results of the expert raters. Finally we will contrast the software outputs and the raters' judgments.

[5] http://www-01.ibm.com/software/analytics/spss/
[6] http://www.r-project.org/

5.1 Software-based voice emotion recognition

Table 1 shows the requested emotions of the participants contrasted with software recognition results. These numbers are taken from all 960 emotions. Each requested emotion is separated in two rows that intersect with the recognized emotions by the software. Our software has the highest recognition rate for the anger expression (83.3 %) and the lowest recognition rate for the surprise expression (54.2 %) (See Table 1 for the confusion matrix between the emotions).

Please note that the obtained differences between software and requested emotions are not necessarily software faults but could also indicate that participants were sometimes unable to mimic the requested emotions. The software had in particular problems to distinguish sad from neutral, neutral from sad, disgust from anger, happy from surprise, happy from anger, and also to distinguish surprise from happy. Error rates are typically between 0.5 and 27 % in all cases in Table 1. The most significant confusions between the emotions are considered in four important groups. The software confused 27 % of the surprise as happy, 25 % of the sad as neutral, 18.7 % of disgust as anger, and 17.6 % of neutral as sad.

The rows from Table 1 show that all seven basic emotions have different distributions for being confused as of the other emotions. In other words, they have different discrimination rates. The results show that the anger expression has the highest recognition rate (83.3 %). It followed by neutral (70.2 %), fear (68.7 %), sad (66.7 %), happy (63.7 %), disgust (62.5 %). The surprise expression has the lowest recognition rate (54.2 %) (Table 1). The surprise is easily confused with happy 27 %, anger 10.7 %, neutral 4.2 %, fear 2.1 %, and disgust 2.1 %, respectively. This result is in accordance

Table 1 Requested emotions and recognized emotions by the software

		Recognized emotion by the software							Total
		Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral	
Requested emotion	Happy	61	0	9	5	6	9	6	96
		63.7 %	0 %	9.3 %	5.2 %	6.25 %	9.3 %	6.25 %	100 %
	Sad	0	32	0	3	1	0	12	48
		0 %	66.7 %	0 %	6.2 %	2.1 %	0 %	25 %	100 %
	Surprise	13	0	26	1	1	5	2	48
		27 %	0 %	54.2 %	2.1 %	2.1 %	10.4 %	4.2 %	100 %
	Fear	2	2	5	33	4	2	0	48
		4.2 %	4.2 %	10.4 %	68.7 %	8.3 %	4.2 %	0 %	100 %
	Disgust	3	1	1	4	30	9	0	48
		6.2 %	2.1 %	2.1 %	8.4 %	62.5 %	18.7 %	0 %	100 %
	Anger	2	0	2	1	3	40	0	48
		4.2 %	0 %	4.2 %	2.1 %	6.2 %	83.3 %	0 %	100 %
	Neutral	23	110	6	22	22	3	438	624
		3.7 %	17.6 %	1 %	3.5 %	3.5 %	0.5 %	70.2 %	100 %
Total		104	145	49	69	67	68	458	960

with Chen and colleagues (Chen et al. 2012), who stated that the best emotion to be recognized is anger and the most difficult emotions to mimic accurately are surprise and happy. The emotion that shows best discrimination from other emotions is sad, even though its rank is placed in the middle of the list of the emotions. Sad is not confused with happy, surprise, and anger at all. 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: 67 % (see Table 1).

According to the raters' analysis results, Table 2 specifies that the participants were able to mimic the requested emotion in 826 occurrences (86 %). In 87 occurrences (9 %) there was sustained disagreement between raters. In 5 % of the cases the raters agreed that participants were unable to mimic requested emotions (47 times). Participants are best at mimicking neutral (94 %), followed by happy (76 %), fear (73 %), surprise (71 %), disgust (68.8 %), anger (68.8 %), and worst at mimicking sad (62.5 %).

Table 3 presents another emotions agreement matrix. This table shows the requested emotions of the participants contrasted with the voice emotion recognition software results, while excluding both the 'unable to mimic' records and the records on which the raters disagreed from the dataset. We therefore, re-calculated the results of each emotion individually and in overall.

The result show positive changes when the 'unable to mimic' emotions and the cases where raters disagree are removed from Table 1. By removing these cases from this table, we importantly have eliminated errors caused by participants' limited acting skills. This is why Table 3 is more informative than Table 1. Accordingly, accuracies of all emotions move toward higher values. For example, happy is changed from 63.7 % to 74 %, sad from 66.7 % to 83.4 %, surprise from 54.2 % to 64.7 %, and anger from 83.3 % to 94 % (compare Tables 1 and 3). 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: 75.7 % (based on Table 3). This result is in accordance with El Ayadi and colleagues (El Ayadi et al. 2011), who stated that the accuracies for existing emotion recognition software solutions are from 51.19 % to 81.94 %.

5.2 Results of the raters for recognizing emotions

Hereafter, we describe how the raters detected participants' emotions from their recorded wav files. The disagreement between the raters, which was 17 % before the

Table 2 Raters' agreements and disagreements about 960 mimicked emotions

	Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral	Total
Raters agree: Able to mimic	73	30	34	35	33	33	588	826
	76 %	62.5 %	71 %	73 %	68.8 %	68.8 %	94 %	86 %
Raters disagree: Able/unable to mimic	16	9	10	7	11	9	25	87
	16.7 %	18.7 %	21 %	14.6 %	23 %	18.7 %	4 %	9 %
Raters agree: Unable to mimic	7	9	4	6	5	6	10	47
	7.3 %	18.8 %	8 %	12.4 %	8.2 %	12.5 %	2 %	5 %
								100 %

Table 3 Requested emotions and emotions recognized by the software – These data (826 emotions) are derived from 960 mimicked emotions, exclusive both 'unable to mimic' records and the records on which the raters disagreed (134 emotions in total)

		Recognized emotion by the software							Total
		Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral	
Requested Emotion	Happy	54	0	4	4	3	6	2	73
		74 %	0 %	5.5 %	5.5 %	4.1 %	8.2 %	2.7 %	100 %
	Sad	0	25	0	1	0	0	4	30
		0 %	83.4 %	0 %	3.3 %	0 %	0 %	13.3 %	100 %
	Surprise	6	0	22	0	1	3	2	34
		17.7 %	0 %	64.7 %	0 %	2.9 %	8.8 %	5.9 %	100 %
	Fear	1	1	2	27	3	1	0	35
		2.8 %	2.8 %	5.7 %	77.3 %	8.6 %	2.8 %	0 %	100 %
	Disgust	2	1	1	2	21	5	0	32
		6.25 %	3.1 %	3.1 %	6.25 %	65.7 %	15.6 %	0 %	100 %
	Anger	0	0	0	1	1	31	0	33
		0 %	0 %	0 %	3 %	3 %	94 %	0 %	100 %
	Neutral	22	102	6	20	20	2	417	589
		3.6 %	17.3 %	1 %	3.4 %	3.4 %	0.3 %	71 %	100 %
Total		85	129	35	55	49	48	425	826

negotiation session, was reduced to 9 % at the end of the negotiation session. In order to determine consistency among raters we performed the cross tabulation between the raters and also interrater reliability analysis using the Kappa statistic approach. We calculated and presented the Kappa value for the original ratings before negotiation. We have 960 displayed emotions (see Table 1) whose recognition is rated and negotiated by two raters as being one of the seven basic emotions. The cross tabulation data (agreement matrix between the raters) are given in Table 4. Each recognized emotion by one rater is separated in two rows that intersect with the recognized emotions by the other rater. The first row indicates the number of occurrences of the recognized emotion and the second row displays the percentage of each recognized emotion.

Cross tabulation analysis between the raters indicates that the neutral expression has the highest agreement (96.2 %). It followed by fear (73.7 %), happy (65.8 %), sad (65.5 %), surprise (61 %), and anger (52 %). The disgust expression has the lowest agreement between them (47.5 %) (Table 4). Our data analysis between the two raters indicate that they have more difficulty in distinguishing between 'surprise and happy', 'disgust and anger', and 'sad and neutral' groups. Indeed, the raters had to correct their recognition rate after the negotiation session mostly between these three groups. Analysing of the Kappa statistic of the Table 4 stresses the agreement among the raters. The result with 95 % confidence among the raters reveals that the interrater reliability of the raters was calculated to be Kappa=0.704 ($p<0.001$). Therefore a substantial agreement among raters is obtained based on Landis and Koch interpretation of Kappa values (Landis and Koch 1977). From the literature we know that the human recognition accuracy was 65 % in (Nwe et al. 2003) and 80 % in (Burkhardt et al. 2005).

Table 4 Rater1 * rater2 cross tabulation – All 960 emotions are rated by both raters

		Recognized emotion by the Rater 2							Total
		Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral	
Recognized emotion by the Rater 1	Happy	71	0	17	1	4	0	15	108
		65.8 %	0 %	15.7 %	0.9 %	3.7 %	0 %	13.9 %	100 %
	Sad	0	38	0	3	1	0	16	58
		0 %	65.5 %	0 %	5.2 %	1.7 %	0 %	27.6 %	100 %
	Surprise	13	0	28	0	2	1	2	46
		28.3 %	0 %	61 %	0 %	4.3 %	2.1 %	4.3 %	100 %
	Fear	0	3	1	28	4	2	0	38
		0 %	7.9 %	2.6 %	73.7 %	10.5 %	5.3 %	0 %	100 %
	Disgust	0	0	6	3	28	18	4	59
		0 %	0 %	10.1 %	5.1 %	47.5 %	30.5 %	6.8 %	100 %
	Anger	0	0	4	3	17	26	0	50
		0 %	0 %	8 %	6 %	34 %	52 %	0 %	100 %
	Neutral	4	12	3	1	3	0	578	601
		0.7 %	2 %	0.5 %	0.1 %	0.5 %	0 %	96.2 %	100 %
Total		88	53	59	39	59	47	615	960

5.3 Results of contrasting the software outputs and the raters' ratings

Using the raters' agreement about the displayed emotions as a reference we report the reliability analysis of our software-based emotion recognition using 95 % confidence intervals and $p<0.001$ in Table 5. It shows the Kappa value of each emotion and the overall Kappa value amongst raters, and the software derived from 826 emotions. This number (826) is used as both raters agreed that the participants were able to mimic the requested emotions (see Table 2).

An analysis of the Kappa values for each emotion reveals that most agreement is for the emotion category of anger (Kappa=0.865, $p<.001$) followed by happy 0.819, surprise 0.810, fear 0.764, neutral 0.756, disgust 0.719, and sad 0.467.

The result with 95 % confidence among the raters and the software reveals that the interrater reliability of them was calculated to be Kappa=0.743 ($p<0.001$). Therefore a substantial agreement among the raters and the software is obtained based on Landis and Koch interpretation of Kappa values (Landis and Koch 1977). We should state that this Kappa value (0.743) is calculated based upon the raters' opinions and the software's results; however the overall accuracy (67 %) of our software is calculated based

Table 5 The overall Kappa of 826 occurrences and the Kappa value of each emotion among raters and the software artifact

Name of emotion	Happy	Sad	Surprise	Fear	Disgust	Anger	Neutral
Kappa value	0.819	0.467	0.810	0.764	0.719	0.865	0.756
Overall Kappa=0.743							

on the uniform distribution of the diagonal in Table 1 and the Kappa value of 0.525 is calculated based upon the requested emotions and the recognized emotions.

5.4 Results of participants' opinion

The Google's questionnaire indicated that seven of twelve participants found that it was easy or somewhat easy for them to mimic the requested emotions in the given tasks (see the difficulty of the given tasks). Table 6 presented the opinion of the participants. Nine out of twelve agreed or mildly agreed that the feedback supported them to lead and mimic the emotions. The feedback also helped them to become more aware of their own emotions. The self-assurance factor indicates that nine of twelve participants completely disagreed, disagreed, or mildly disagreed that they were able to mimic the requested emotions in the given tasks. This factor indicates the necessity of this study that most probably can help them training their acting skills as well as improving their communication skills.

All participants agreed that the instructions for the given tasks were clear to them to perform the tasks. All the tasks were interesting or mildly interesting for the participants to perform. They indicated no distraction during performance. Except two participants it was easy to understand that the participants did not regard themselves as actors or they don't have any clear idea about this skill.

6 Ethical implications

We follow a restricted and protected data approach for our learning analytics in this study. The users' privacy, including making the current and the future data of the

Table 6 Opinion of the participants

	Answers by the participants								
			1	2	3	4	5	6	7	Total
Questions	Difficulty	It was easy for me to mimic the requested emotions in the given tasks	–	2	2	1	3	4	–	12
	Feedback	The feedback did help me to mimic the emotions in the given tasks	–	–	1	2	3	6	–	
	Self-assurance	I am confident that I was able to mimic the requested emotions in the given tasks	1	1	7	1	1	1	–	
	Instructiveness	The instructions for the given tasks were clear to me	–	–	–	–	4	7	1	
	Attractiveness	The given tasks were interesting	–	–	–	–	8	4	–	
	Concentration	I could easily focus on the given tasks and was not distracted by other factors	–	–	–	1	1	6	4	
	Acting skills	I regard myself as a good actor	–	4	2	4	–	2	–	

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

participants available to public without their prior permission, are serious issues and we are aware of the consequences.

7 Discussion

This study contrasted requested emotions of participants with software recognition results from the voice emotion recognition part of FILTWAM. Two human raters contributed as expert observers judging the experimental results. This study showed a substantial agreement between the raters and the software with an overall Kappa value of 0.743; only including cases of full agreement between human raters (826 emotions were considered). The Kappa value of 0.743 indicates that the software reliably and accurately establishes the users' emotions. We used G*Power tool (Erdfelder et al. 1996) and applied T tests and F tests. We calculated the actual power = 0.95 of our dataset with error probability = 0.05 and effect size = 0.12 and realized that the best total sample size required for our study is between 741 and 904 occurrences. We used 960 occurrences for sampling the 'requested emotions', thus this criteria was met.

The best recognized emotion was anger, 94 %, followed by sadness, 83.4 %, fear, 77.3 %, happiness, 74 %, neutral 71 %, disgust, 65.7 %, and surprise, 64.7 %. From the voice power perspective, the result shows that the dominate emotion (anger) and a less dominate emotion (sadness) are ranked higher than other emotions. In the 134 cases where one or both raters indicated that the participants were unable to mimic emotions, the participants had problems mimicking neutral in 35 cases followed by happiness 23 cases, sadness 18 cases, disgust 16 cases, anger 15 cases, surprise 14 cases, and fear 13 cases. This is largely in agreement with Chen and colleagues (Chen et al. 2012), who found that the most difficult voice emotion to mimic accurately is happiness and the easiest one is anger. Moreover, our analysis also confirms Chen and colleagues (Chen et al. 2012) finding stating that the two sets of emotions – happiness/surprise and anger/disgust – are difficult to distinguish from each other and are often wrongly classified. The overall accuracy of our software based on the requested emotions and the recognized emotions is 67 %. It is worth noting that the software is incapable of checking the extra category of 'not being able to mimic' that was reported by the two experts, meaning the software has an inherently lower accuracy.

We found three sets of emotions that were difficult to distinguish: anger/disgust, happiness/surprise, and neutral/sad. As reported elsewhere (Burkhardt et al. 2005; Chen et al. 2012; Nwe et al. 2003), these sets of emotions are consistently confusing emotions that are difficult to distinguish. This is likely the reason why there are three commonly confused sets of emotions in Tables 1 and 3. We invited non-actors to participate in order to avoid extreme emotional expressions that normally occur in actors performances. Krahmer and Swerts have shown that using actors, although they evidently have better acting skills than layman, does not lead to more realistic expressions (i.e., authentic, spontaneous) (Krahmer and Swerts 2011). However, as youngsters and older adults are not equally good in mimicking different basic emotions (e.g., older adults are less good in mimicking sadness and happiness than youngsters, but older adults mimic disgust better than youngsters), it is acknowledged that the sample of participants may influence the accuracy of the software (Huhnel et al. 2014). Our participants were middle-aged adults. It is possible that this sample of middle-aged

adults mitigates the strength and weaknesses of both older adults and youngsters, but this has not yet been investigated. No gender differences in mimicry for both younger male and female participants have been reported (Huhnel et al. 2014), but it is possible that gender differences exist in older age groups. El Ayadi and colleagues (El Ayadi et al. 2011) report accuracies for existing expression recognition software solutions ranging from 51.19 % to 81.94 %, based on the offline classification approaches and algorithms. But our software's results are based on the online classification approaches for emotion recognition. In 134 cases (13.95 %) our participants were unable to mimic the requested emotions, but all participants found the software easy and straightforward to use. We fulfilled our basic requirements for 1) an unobtrusive approach, 2) with an objective method 3) with inexpensive and ubiquitous equipment (microphone), and 4) offering a real-time software artefact with easy to use interface. However, there are still a number of limitations of the study that require further investigation. Uncertainty in the detection of a specific emotion remains open. Perhaps this could be overcome by applying multimodal sources for emotion detection. Our database is currently a language-dependent database for English speakers, but it could be extended to other databases for other languages. Nonetheless, certain issues remain open with regard to the participants' characteristics and languages that should be further investigated. For example, dealing with quickly changing emotions or multiple emotions occurring simultaneously are challenges that our software artefact is not yet prepared to handle.

8 Summary and outlook

Our previous research on automatic emotion recognition from facial expressions has shown that it is possible to measure emotions from a face emotion recognition software artefact with sufficient reliability in real-time (Bahreini et al. 2014). In this study, which is an extension of our previous study (Bahreini et al. 2013), we built on automatic emotion recognition from vocal expressions and investigated the suitability of a simple microphone for continuously and unobtrusively gathering affective user data in an e-learning context.

It appears that the rate of the affective computing tool for emotion recognition can be further improved by combining the voice emotion recognition software artefact with the face emotion recognition software artefact of FILTWAM. This would offer an innovative approach for applying emotion recognition in affective e-learning (Bahreini et al. 2014; Sebe 2009). A study by Sebe and colleagues showed that the average person-dependent emotion recognition accuracy is significantly enhanced when both visual and acoustic information are used in classification (Sebe et al. 2006). The average recognition accuracy is about 56 % for the face-only classifier, about 45 % for the prosody-only classifier, but around 90 % for the multimodal classifier. This suggests that combining multimodal data for inferring emotions, and explains why our future study will combine face and voice expressions when triggering support during training in e-learning settings. Contemporary research on affect recognition also focuses on approaches that can handle visual and acoustic recordings of affective states (Zeng et al. 2009). Effectively, the FILTWAM framework is designed for encompassing multimodal data.

9 Conclusion

This paper described real-time speech emotion recognition for affective learning covered by the FILTWAM framework. The approach aims to continuously and unobtrusively monitor learners' behaviour during e-learning and to interpret this input into emotional states. FILTWAM aims to improve learning when using webcams and microphones as input devices and exploits multimodal emotion recognition of learners during e-learning, linking emotion detection to adapted learning activities. We continue Sebe's (2009) approach of combining both visual and audio information for classification to improve the accuracy of detecting basic emotions. FILTWAM anticipates the increased importance of affective user states and cognitive states in gamified pedagogical scaffolding. Our approach supports the usage of ubiquitous consumer equipment, which is portable and easy to use. Our study has demonstrated that learners are able to improve their communication skills when using this approach. They are able to become more aware of their own emotions during both good news and bad news conversations. Our software feedback supports this awareness. Although we have considered only seven basic emotions in this study, our software can be extended to include additional emotions. The outcomes of FILTWAM can influence different groups' interests in virtual settings. The integration of the voice emotion recognition and the face emotion recognition software artefacts, and processing these two artefacts in an online affective gamified e-learning environment are future steps for achieving FILTWAM's full potential for e-learning.

Acknowledgments We thank our colleagues at Welten Institute of the Open University Netherlands who participated in this study. We likewise thank the two raters who helped us to rate the recorded files. We also thank the Netherlands Laboratory for Lifelong Learning (NELLL) of the Open University Netherlands that has sponsored this research.

References

Bachiller, C., Hernandez, C., & Sastre, J. (2010). Collaborative learning, research and science promotion in a multidisciplinary scenario: information and communications technology and music. Proceedings of the International Conference on Engineering Education (pp. 1–8). Gliwice, Poland.

Bahreini, K., Nadolski, R., Qi, W., & Westera, W. (2012). 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)* (pp. 39–48). Cork: Academic Publishing International Limited Reading.

Bahreini, K., Nadolski, R., & Westera, W. (2013). FLITWAM and Voice Emotion Recognition. Games and Learning Alliance (GaLA) Conference. Paris, France, 23–25.

Bahreini, K., Nadolski, R., & Westera, W. (2014). Towards Multimodal Emotion Recognition in E-learning Environments. Interactive Learning Environments 1–16.

Beale, R., & Creed, C. (2009). Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies, 67*(9), 755–776.

Batliner, A., Fischer, K., Hubera, R., Spilkera, J., & Nöth, E. (2003). How to find trouble in communication. *Speech Communication, 40*, 117–143.

Ben Ammar, M., Neji, M., Alimi, A. M., & Gouardères, G. (2010). The affective tutoring system. *Expert Systems with Applications, 37*(4), 3013–3023.

Bozkurt, E., Erzin, E. Erdem, Ç.E., & Erdem, A.T. (2009). Improving automatic emotion recognition for speech signals. Accessible on: http://www.isca-speech.org/archive/interspeech_2009/i09_0324.html. Last Accessed 7 July 2014.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In proceedings of the Inter speech Lissabon, Portugal. 1517–1520.

Chen, L., Mao, X., Xue, Y., & Cheng, L. L. (2012). Speech emotion recognition: features and classification models. *Digital Signal Processing, 22*(6), 1154–1160.

Chibelushi, C.C., & Bourel, F. (2003). Facial expression recognition: a brief tutorial overview. Available Online in Compendium of Computer Vision.

Ebner, M. (2007). E-Learning 2.0 = e-Learning 1.0 + Web 2.0?. The Second International Conference on Availability, Reliability and Security (ARES), 1235–1239.

Ekman, P., & Friesen, W.V. (1978). *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: a general power analysis program. *Behavior Research Methods, Instruments, and Computers., 28*, 1–11.

Feidakis, M., Daradoumis, T., & Caballe, S. (2011). Emotion Measurement in Intelligent Tutoring Systems: What, When and How to Measure. Third International Conference on Intelligent Networking and Collaborative Systems, 807–812.

Happy, S. L. Dasgupta, A., Patnaik, P., Routray, A. (2013). Automated Alertness and Emotion Detection for Empathic Feedback during e-Learning. IEEE Fifth International Conference on Technology for Education (T4E). 47–50.

Huhnel, I., Fölster, M., Werheid, K., & Hess, U. (2014). Empathic reactions of younger and older adults: no age related decline in affective responding. *Journal of Experimental Social Psychology, 50*, 136–143.

Jianhua, T., Tieniu, T., & RosalindW, P. (2005). Affective computing: a review. Affective computing and intelligent interaction. *Springer Berlin Heidelberg, 3784*, 981–995.

Jones, C., & Sutherland, J. (2008). Acoustic emotion recognition for affective computer gaming. In C. Peter & R. Beale (Eds.), *Affect and Emotion in Human-Computer Interaction. LNCS. 4868*. Heidelberg: Springer.

Krahmer, E., & Swerts, M. (2011). Audiovisual expression of emotions in communication. Philips research book series. *Springer Netherlands, 12*, 85–106.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lang, G., & van der Molen, H. T. (2008). *Psychologische gespreksvoering book*. The Netherlands: Open University of the Netherlands, Heerlen.

Liu, X., Zhang, L., Yadegar, J., & Kamat, N. (2011). A Robust Multi-Modal Emotion Recognition Framework for Intelligent Tutoring Systems. Advanced Learning Technologies, IEEE International Conference on Advanced Learning Technologies. 63–65.

López-Cózar, R., Silovsky, J., & Kroul, M. (2011). Enhancement of emotion detection in spoken dialogue systems by combining several information sources. *Speech Communication, 53*(9–10), 1210–1228.

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. Interspeech ICSLP, 17–21, Pittsburgh, Pennsylvania.

Nwe, T., Foo, S., & De Silva, L. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*, 603–623.

Pekrun, R. (1992). The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. *Journal of Applied Psychology, 41*, 359–376.

Pfister, T., & Robinson, P. (2011). Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Transactions on Affective Computing, 2*(2), 66–78.

Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., & Bigdeli, A. (2008). How do you know that I don't understand?" A look at the future of intelligent tutoring systems. *Computers in Human Behavior, 24*(4), 1342–1363.

Schröder, M. (2009). The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems. Advances in Human Computer Interaction. 2010 (319406), Hindawi Publishing Coorporation.

Sebe, N., Cohen, I., Gevers, T., & Huang, T.S. (2006). Emotion recognition based on joint visual and audio cues. 18th International Conference on Pattern recognition. 1136–1139. Hong Kong.

Sebe, N. (2009). Multimodal interfaces: challenges and perspectives. *Journal of Ambient Intelligence and Smart Environments, 1*(1), 23–30.

Van der Molen, H. T., & Gramsbergen-Hoogland, Y. H. (2005). *Communication in Organizations: Basic Skills and Conversation Models*. New York: Psychology Press.

Vogt, T. André, & E. Bee, N. (2008a). EmoVoice - A framework for online recognition of emotions from voice. In proceedings of workshop on Perception and Interactive Technologies for Speech-Based Systems.

Vogt, T., Andre, E., & Wagner, J. (2008b). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization. *LNCS, 4868*, 75–91.

Wagner, J., Lingenfelser, F., & Andre, E. (2011). The Social Signal Interpretation Framework (SSI) for Real Time Signal Processing and Recognitions, In Proceedings of INTERSPEECH, Florence, Italy, 2011.

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & Andre, E. (2013). The Social Signal Interpretation (SSI) Framework: Multimodal Signal Processing and Recognition in Real-time. Proceedings of the 21st ACM International Conference on Multimedia, MM '13. Barcelona, Spain. 831–834.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58.