Introducing an Emotion-Driven Assistance System for Cognitively Impaired Individuals

Simone Hantke[1,2(✉)], Christian Cohrs[3], Maximilian Schmitt[1], Benjamin Tannert[3], Florian Lütkebohmert[3], Mathias Detmers[3], Heidi Schelhowe[3], and Björn Schuller[1,4]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany
simone.hantke@informatik.uni-augsburg.de
[2] Machine Intelligence and Signal Processing Group, Technische Universität München, Munich, Germany
[3] Working Group on Digital Media in Education, University of Bremen, Bremen, Germany
[4] GLAM – Group on Language, Audio and Music, Imperial College London, London, UK

Abstract. Mental, neurological and/or physical disabilities often affect individuals' cognitive processes, which in turn can introduce difficulties with remembering what they have learnt. Therefore, completing trivial daily tasks can be challenging and supervision or help from others is constantly needed. In this regard, these individuals with special needs can benefit from nowadays advanced assistance techniques. Within this contribution, a language-driven, workplace integrated, assistance system is being proposed, supporting disabled individuals in the handling of certain activities while taking into account their emotional-cognitive constitution and state. In this context, we present a set of baseline results for emotion recognition tasks and conduct machine learning experiments to benchmark the performance of an automatic emotion recognition system on the collected data. We show that this is a challenging task that can nevertheless be tackled with state-of-the-art methodologies.

Keywords: Speech-driven assistive technology · Disabilities · Affect Speech and emotion recognition

1 Introduction

Technology is growing rapidly and opens a new world of opportunities in different areas of health care. In this context, a wide range of applications for many health conditions such as dementia [26], depression [2], or Parkinson's disease [28] are being developed. Nowadays mobile communications and network technologies for health care entail the usage of portable devices with the capability to create,

K. Miesenberger and G. Kouroupetroglou (Eds.): ICCHP 2018, LNCS 10896, pp. 486–494, 2018.
https://doi.org/10.1007/978-3-319-94277-3_75

analyse, store, retrieve, and transmit data in real-time between the users, for the purpose of improving individuals' safety and quality of life [9].

As well as being applied to improve life quality, these systems are capable of facilitating the communication between clinicians and patients [28]. In this regard, remote monitoring systems have amongst others been proposed for asthma patients [15], the tracking of patients with dementia [14], or to support treatment of sleep apnoea [8]. Such assistance systems can also be adapted for unobtrusively recognising stress from human voice [13], to perform suicide prevention [12], or to enable individuals with special needs to both join the workforce and aid them at their job [24].

This contribution focuses on people with cognitive impairments which need special ways to learn new things and keep them in mind. Especially at their workplace, there is necessity to recall working steps in mind to fulfil the work without becoming injured or to avoid inappropriate actions. For these individuals it is often hard to learn new tasks, as they are not able to abstract the work process. Therefore, a special explanation right at the machine or workstation is needed [23]. Another characteristic of this target group is their weak ability in staying focused [23]. Often, small things like music or statements of a colleague can distract these individuals from their work-task. Furthermore, the ability to stay focused or other abilities differ from their normal behaviour according to their emotional status and their health constitution [11].

Psychological research results on people with cognitive disabilities indicate a relation between the voice and articulation on the one hand, and the emotional status on the other hand [11,25]. In this regard – within the German national ***Emot****ional sensitivity* ***As****sistance* ***S****ystem for people with disabilities* (*EmotAsS*) project – a language-driven, workplace integrated, assistance system is being developed. The first-of-its kind system aims at supporting individuals with mental, neurological, and/or physical disabilities in the handling of certain activities while taking into account their emotional-cognitive constitution and state.

2 Exemplary Application Area

The first application of the proposed system will be located at the working shelter with disabled employees working in a cleaning department. The complexity of the given tasks within this department lies in the nature of having the knowledge – and more importantly remembering – the different usages and compositions of the cleaning devices for the different tasks to perform, e. g., cleaning the workshop areas, the bathrooms, or the offices. Therefore, the cleaning trolley needs to be prepared with its special equipment depending on these different tasks. For the employees, the proposed system can make it easier to learn and remember how to equip the trolley correctly.

We imagine the following situation: Alex is an employee of the cleaning department and was given the task to clean the wood workshop. Alex is unsure about the required supplies and embarrassed to ask the supervisor again. Fortunately, the proposed assistance system can provide the answer, so Alex asks the speech-driven system instead. The integrated emotion recognition component

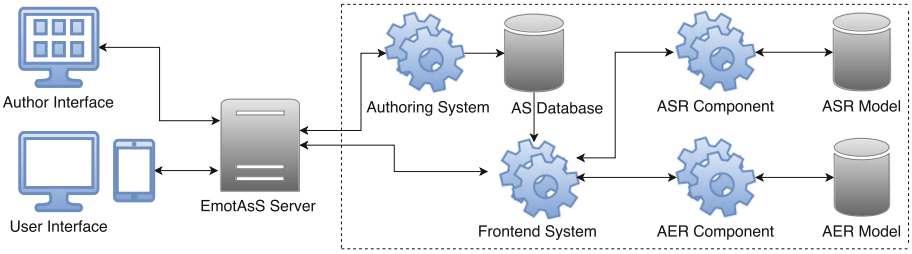

Fig. 1. Overview of the hardware and software components of the assistance system (AS), including the automatic speech recognition (ASR) and automatic emotion recognition (AER) component.

detects that Alex seems to be in a bad mood. In this regard, the system will present longer and more detailed descriptions on the task Alex likes to perform. After a few shown working steps the system detects that Alex' mood has brightened and now presents slightly shorter and less detailed descriptions how to equip the trolley. With the help of the assistance system, Alex was able to perform this task without help of the supervisor, which facilitate independent work, and thus strengthen self-confidence.

3 The System

The system is composed of several hardware elements, consisting of a central server-system and several thin-clients. The applied software elements are divided into a web-application, which provides the assistance and authoring frontends to the end users (here: disabled employees of the sheltered workplace) and authors (here: care-taking personal), and the automatic speech and emotion recognition sub-systems. The authoring and frontend system was implemented as a web application to enable the system to be accessible from several devices such as notebooks and tablets, granting the users mobility during daily use, as well as easy integration into daily work-life. Speech and emotion recognition is processed by a powerful multi-core i7-System, which runs the webserver for the authoring and frontend as well. Clients access the system via web browser. Speech data and recognition results are provided to clients and server via TCP sockets and web-sockets. To be agile as well as structured in its development, the system was implemented via the web application framework Ruby on Rails.

The authoring interface acts as the management system (cf. Fig. 1) and allows creating and accessing user profiles of both end users and authors, instructions for work steps and devices, adding or editing interventions, and accessing frontend statistics. Finally, all data is stored in a database. The assistance frontend is the 'face' of the system, offering instructions, records speech of the end users and passes it to the *automatic speech recognition* (ASR) and *automatic emotion recognition* (AER) components. ASR provides verbal content detected in speech and the assistance system compares the content with commands specified in the database. In addition, the frontend system also offers control inputs via a

specialised keyboard without the need for speech input. This allows users with speech disabilities to use the assistance system to a certain degree. Simultaneously, AER is processing the speech recordings of a user, detecting *arousal* (describing how strong or weak an emotion is) and *valence* (describing how positive or negative an emotion is) within the speech sample. Depending on the emotion recognised, the frontend will display emotion-specific content to the user.

4 Emotional Speech Data Collection

To develop the assistance system, emotional speech data from disabled employees were recorded at a workplace shelter as described in detail in [7]. Seventeen participants (ten female and seven male, ages range from 19 to 58 years with a mean age of 31.6 years, and standard deviation of 11.7 years) agreed to take part in the experiment and provided data relating to their personal and health issues including their form of disability. As there are strict ethic restrictions on the data, no further details on the disability of the subjects can be given, but can be clustered into mental, neurological, and physical disabilities: thirteen participants are mentally disabled, three neurological, and one has multiple disabilities.

Taking into account the daily or even hourly mood changes of the participants, a special recording set-up was developed, as not to add undue stress. To achieve a high, but also realistic audio quality, the recordings took place in a working room with equal set-up and conditions for each recording session. The participants had to sit down in front of the recording equipments. An experimental supervisor, an internal occupational therapist, and an internal psychologist of the shelter were sitting next to the participants all the time to communicate and help them through the given tasks. The recorded data consists of spontaneous speech and was recorded by giving the participants different tasks related to different contents [7]. In this way, questions were raised about professional life and certain tasks to be accomplished. The tasks were designed in such a way that the mood of the participants can be shown and their emotions are being provoked. To ensure a professional management of possibly expressed emotions, the tasks were performed in supervision of a psychologist.

The data was annotated by 29 annotators using the crowdsourcing platform iHEARu-PLAY [6]. Labels were gathered giving the annotators the choice to select from the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise), as well as 'neutral' [7]. Moreover, each utterance was annotated on a 5-point likert scale, to represent the intelligibility of the speech, so the data can be used for automatic speech recognition tasks.

5 Automatic Speech and Emotion Recognition

For both the ASR and AER subsystems, the recorded voice is first chunked on the client side, using the *WebRTC Voice Activity Detector*[1]. The resulting

[1] https://pypi.python.org/pypi/webrtcvad.

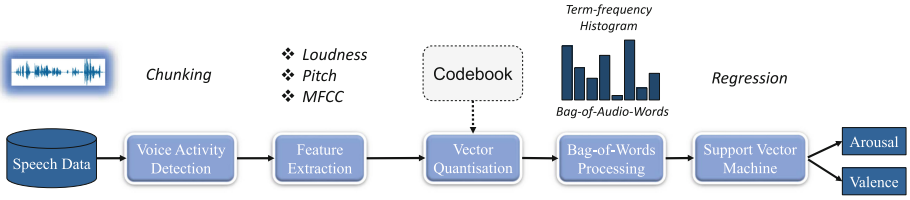

Fig. 2. Schematic overview of the automatic emotion recognition system.

chunks of a duration of approximately 2 to 5 s are then sent to the server via a TCP socket.

The ASR subsystem was integrated using the open-source toolkit KALDI [16]. MFCCs as the most commonly used speech features were extracted and normalised on chunk level. The *acoustic model* consists of a state-of-the-art *max-out* Deep Neural Network [29] and was trained on 160 h of German broadcast news [27]. As a *language model*, a *4-gram Kneser-Ney backoff language model* is used [10], trained on 307 million running words from German newspaper articles. A grapheme-to-phoneme conversion is used to generate a pronunciation model for words not present in the training data of the acoustic model. Evaluation on the *NoXi* database of video chat recordings [1] shows a *word error rate* of approximately 30%.

The AER subsystem processes the incoming speech in the way as illustrated in Fig. 2: first, 65 acoustic features are extracted from small blocks (frames) of the chunk. The features are defined in the COMPARE feature set [22] and include prosodic features, such as *loudness* and *pitch*, *spectral* descriptors, and *Mel-frequency cepstral coefficients* (MFCC), which are a feature widely-used in speech analysis [21]. Then, the acoustic features of one chunk are summarised using the *bag-of-audio-words* (BoAW) approach, which has been proven to be very suitable for the task at hand [19]. In this method, the low-level descriptors (LLDs) from each frame are quantised using a 'codebook' of audio words, previously learnt from data of the same domain using clustering. After this step of *vector quantisation*, the frequencies of each audio word within the whole utterance are counted, resulting in a so-called *term-frequency histogram* or BoAW. For the proposed AER system, we used a codebook of size 1 000 and 10 assignments per frame. This fixed-length histogram is then the input to a *support vector regression*, predicting continuous values for the emotional dimensions arousal and valence. Using the in-the-wild database SEWA of naturalistic and spontaneous humans emotions [17], *concordance correlation coefficients* of .47 and .43 are achieved for arousal and valence, respectively [20]. The implementation of the whole AER subsystem uses only the open-source tools OPENSMILE [4], OPENXBOW [20], and LIBLINEAR [5].

Evaluating the proposed system, extensive machine learning experiments were run on the collected database. As the basic emotions disgust, fear, and surprise are sparse, only the categories anger, happiness, and sadness were

Table 1. Class distribution and classification results on the EmotAsS database (*A(nger)*, *H(appiness)*, *N(eutral)*, *S(adness)*) in terms of unweighted average recall (UAR) [%]; C: complexity of the SVM, CS: codebook size.

Class distribution					ComParE + SVM			BoAW + SVM			Fusion
	Train.	Devel.	Test	Σ	C	Devel.	Test	CS	Devel.	Test	Test
A	125	50	272	447	10^{-6}	34.2	41.3	250	38.7	36.4	43.4
H	743	965	650	2 358	10^{-5}	37.8	43.1	500	40.5	36.5	
N	2 287	2 842	2 024	7 153	10^{-4}	28.2	38.4	1 000	39.8	38.1	
S	187	329	153	669	10^{-3}	29.9	35.4	2 000	38.1	41.3	
Σ	3 342	4 186	3 099	10 627	10^{-2}	29.9	33.1	4 000	37.9	39.2	

considered for the performed experiments, together with a neutral background class. For the experiments – also employed as a baseline in the Interspeech 2018 ComParE Challenge [22] – we split the data into three partitions, training, development, and test partition, with five subjects in each split. Both the ComParE acoustic feature set consisting of 6 373 *supra-segmental* acoustic features and BoAW representations of the corresponding 130 LLDs were considered. A support vector machine (SVM) was used as a classifier. Both the complexity of the SVM and the codebook size for the BoAW approach were optimised; the model was re-trained on the fusion of the training and the development set to receive the final performance estimate on the test set. Results in terms of the *unweighted average recall (UAR)* are given in Table 1 for different hyperparameters. For BoAW, the complexity has been optimised on the development set for the shown configurations. Moreover, a late fusion of the predictions of both models was evaluated, considering for each data instance the prediction with the larger confidence. Fusing the outputs of the two models performing best on the test set gives a UAR of 43.4%, slightly outperforming the single models. These results show both that AER for the atypical speech data is a quite challenging task but that it is yet feasible with a potential for improvement.

6 Conclusions and Outlook

A novel speech-driven, emotionally sensitive assistance system has been introduced to support mentally, neurologically and/or physically disabled people at their workplace. Successful baseline experiments evaluating the automatic emotion recognition component have been performed in earlier work [3,7,18] and were herein expanded by performing novel BoAW experiments, leading to a UAR of 43.3% (4 emotion classes) on the collected data and promising a large margin for improvement considering recent machine learning techniques such as *generative models* and *transfer learning*. Furthermore, primary usability evaluations with the target group are currently ongoing, giving first insights into

the promising success of the system. Future work will take into account recent machine learning techniques and focus on comparing the gathered speech emotion recognition results with facial emotion expression observations.

Acknowledgement. The research leading to these results has received funding from the German national BMBF IKT2020-Grant under grant agreement No. 16SV7213 (EmotAsS). We thank all iHEARu-PLAY users for donating their annotations.

References

1. Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres, M.T., Pelachaud, C., André, E., Valstar, M.: The NoXi database: multimodal recordings of mediated novice-expert interactions. In: Proceedings of International Conference on Multimodal Interaction, Glasgow, Scotland, pp. 350–359 (2017)
2. Cummins, N., Vlasenko, B., Sagha, H., Schuller, B.: Enhancing speech-based depression detection through gender dependent vowel-level formant. In: Proceedings of Conference on Artificial Intelligence in Medicine, Stockholm, Sweden, pp. 3266–3270 (2017)
3. Deng, J., Xu, X., Zhang, Z., Frühholz, S., Grandjean, D., Schuller, B.: Fisher kernels on phase-based features for speech emotion recognition. In: Jokinen, K., Wilcock, G. (eds.) Dialogues with Social Robots. LNEE, vol. 999, pp. 195–203. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-2585-3_15
4. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of International Conference on Multimedia, Barcelona, Spain, pp. 835–838 (2013)
5. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
6. Hantke, S., Eyben, F., Appel, T., Schuller, B.: iHEARu-PLAY: introducing a game for crowdsourced data collection for affective computing. In: Proceedings of International Workshop on Automatic Sentiment Analysis in the Wild, Satellite of Conference on Affective Computing and Intelligent Interaction, Xi'an, China, pp. 891–897 (2015)
7. Hantke, S., Sagha, H., Cummins, N., Schuller, B.: Emotional speech of mentally and physically disabled individuals: introducing the EmotAsS database and first findings. In: Proceedings of INTERSPEECH, Stockholm, Sweden, pp. 3137–3141 (2017)
8. Isetta, V., Torres, M., González, K., Ruiz, C., Dalmases, M., Embid, C., Navajas, D., Farré, R., Montserrat, J.M.: A new mHealth application to support treatment of sleep apnoea patients. J. telemedicine and telecare **10**, 14–18 (2015)
9. Istepanian, R., Laxminarayan, S., Pattichis, C.S.: M-Health. Springer, Heidelberg (2006). https://doi.org/10.1007/b137697
10. Kneser, R., Ney, H.: Improved backing-off for M-gram language modeling. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, Detroit, USA, pp. 181–184 (1995)
11. Krannich, D., Zare, S.: Concept and design of a mobile learning support system for mentally disabled people at workplace. In: Proceedings of International Conference on E-Learning in the Workplace, New York, USA, pp. 1–6 (2009)

12. Larsen, M.E., Cummins, N., Boonstra, T.W., O'Dea, B., Tighe, J., Nicholas, J., Shand, F., Epps, J., Christensen, H.: The use of technology in suicide prevention. In: Proceedings of International Conference on Engineering in Medicine and Biology Society, Milan, Italy, pp. 7316–7319 (2015)
13. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Gatica-Perez, D., Choudhury, T.: Stresssense: detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of Conference on Ubiquitous Computing, Pittsburgh, USA, pp. 351–360 (2012)
14. Miskelly, F.: Electronic tracking of patients with dementia and wandering using mobile phone technology. Age Ageing **34**, 497–498 (2005)
15. Namazova-Baranova, L.S., Molodchenkov, A.I., Vishneva, E.A., Antonova, E.V., Smirnov, V.I.: Remote monitoring of children with asthma, being treated in multidisciplinary hospital. In: Proceedings of International Conference on Biomedical Engineering and Computational Technologies, Novosibirsk, Russia, pp. 7–12 (2015)
16. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: Proceedings of International Workshop on Automatic Speech Recognition and Understanding, Hawaii, USA, 4 p (2011)
17. Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M.: Avec 2017: real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, USA, pp. 3–9 (2017)
18. Sagha, H., Deng, J., Gavryukova, M., Han, J., Schuller, B.: Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Shanghai, P. R. China, pp. 5800–5804 (2016)
19. Schmitt, M., Ringeval, F., Schuller, B.: At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech. In: Proceedings of INTERSPEECH, San Francisco, USA, pp. 495–499 (2016)
20. Schmitt, M., Schuller, B.: openXBOW-introducing the passau open-source crossmodal bag-of-words toolkit. J. Mach. Learn. Res. **18**, 1–5 (2017)
21. Schuller, B.: Intelligent Audio Analysis. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36806-6
22. Schuller, B.W., Steidl, S., Batliner, A., Marschik, P.B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F., Rathner, E.M., Bartl-Pokorny, K.D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., Zafeiriou, S.: The INTERSPEECH 2018 computational paralinguistics challenge: atypical & self-assessed affect, crying & heart beats. In: Proceedings of INTERSPEECH, Hyderabad, India, 5 p (2018, to appear)
23. Thiel, O.: Das Familienhandbuch des Staatsinstituts fúr Frúhpádagogik - Lernschwierigkeiten (2010)
24. Verbrugge, L.M., Sevak, P.: Use, type, and efficacy of assistance for disability. J. Gerontol. Ser. B: Psychol. Sci. Soc. Sci. **57**, 366–379 (2002)
25. Vogt, T.: Real-time automatic emotion recognition from speech. Ph.D. thesis, University of Bielefeld (2010)
26. Vuong, N.K., Chan, S., Lau, C.T.: mHealth sensors, techniques, and applications for managing wandering behavior of people with dementia: a review. In: Adibi, S. (ed.) Mobile Health. SSB, vol. 5, pp. 11–42. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-12817-7_2

27. Weninger, F., Schuller, B., Eyben, F., Wöllmer, M., Rigoll, G.: A broadcast news corpus for evaluation and tuning of German LVCSR systems. arXiv.org arXiv:1412.4616, 4 p. (2014)
28. Zapata, B.C., Fernández-Alemán, J.L., Idri, A., Toval, A.: Empirical studies on usability of mHealth apps: a systematic literature review. J. Med. Syst. **39**, 1 (2015)
29. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, pp. 215–219 (2014)