INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Audio-Visual Recognition of Emotional Engagement of People with Dementia

Lars Steinert, Felix Putze, Dennis Küster, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany
lars.steinert@uni-bremen.de

Abstract

Dementia places an immeasurable burden on affected individuals and caregivers. In addition to general cognitive decline, dementia has a negative impact on communication. Technical activation systems are thus in high demand, as cognitive activation may help to moderate the decline. However, effective activation requires sustained engagement - which, in turn, first needs to be reliably recognized. In this study, we examine emotional engagement recognition for People with Dementia (PwD) using non-intrusive biosignals resulting from speech communication and facial expressions. PwD suffering from mild to severe dementia used a tablet-based activation system over multiple sessions. We demonstrate that they retained their ability to verbally express emotional engagement even at severe stages of the disease. For recognition of emotional engagement, we propose an architecture of Bidirectional Long-Short-Term-Memory Networks that combines video information with up to three speech-based feature sets (eGeMAPS, ComParE'13, DeepSpectrum). Using data of 24 PwD, we show that adding speech improves recognition performance significantly compared to a video-only model. Interestingly, disease-progression did not appear to have a substantial impact on recognition performance in this sample. We further discuss the opportunities and challenges of detecting emotional engagement from speech in PwD.
Index Terms: engagement, dementia, activation system

1. Introduction

Roughly 50 million people worldwide are currently suffering from dementia, and this number is expected to triple by 2050 [1]. Secondary therapy involving physical, social and cognitive activation have been shown to positively impact cognitive functioning [2, 3] and can help prevent the magnification of apathy, boredom, depression and loneliness associated with dementia [4]. We follow Cohen's [4] argument that activation has to produce engagement to take effect and adopt the definition of engagement as "the act of being occupied or involved with an external stimulus". Accordingly, a technical system which supports activation of People with Dementia (PwD) requires a means to automatically recognize if users are sufficiently engaged. This study examines if speech can be used to improve video-based recognition of engagement of PwD despite the fact that the speech of those affected may be compromised [5, 6]. We analyze the dataset collected in a previous study comprising 24 PwD who used a technical activation system in an unconstrained care setting. Firstly, we examine to what extent PwD provide emotional and verbal responses throughout all stages of the disease, which is the foundation for the automatic recognition of emotional engagement. Secondly, we investigate if speech can improve recognition performance compared to a video-only model based on an architecture of Bidirectional Long-Short-Term-Memory Networks (BiLSTMs). For this, we use three audio feature sets, namely the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [7], the 2013 Interspeech Computational Paralinguistics Challenge features set (ComParE) [8] and DeepSpectrum features [9] extracted using the pre-trained Convolutional Neural Network (CNN) AlexNet. Lastly, we investigate which factors the classification performance is dependent on. To the best of our knowledge, there has been no prior work investigating the usefulness of speech to improve video-based engagement recognition for a diverse, real-world dataset of PwD.

2. Related Work

The recognition of emotional states using speech is a highly active research area [10]. Whilst a plethora of work has studied the recognition of emotion in healthy individuals [11], much less is known about the practicality of speech for assessing emotional states of PwD. One reason could be disease-related changes of speech, e.g. impairments in the production of prosody [5, 6]. Consequently, speech appears to play an important role in dementia detection. Recent work has shown that speech data can be utilized to detect and predict dementia, even before clinical screening tests can diagnose the disease [12]. Speech has further been investigated in other healthcare contexts such as for the detection of depression [13] or autism spectrum disorder [14]. Within the annual Interspeech ComParE challenge, related research fields have been investigated. One sub-challenge focused on recognizing emotions of elderly individuals [15]. Nazareth et al. [16] demonstrated that lexical and acoustic features can be used to predict emotional valence in spontaneous speech of elderly [16]. Ma et al. [17] introduced a multimodal dataset of spontaneous emotional responses of healthy elders which was collected from YouTube. In a first study, we showed that emotional engagement for PwD can be recognized based on visual and contextual data [18].

3. Data and Annotation

The data used in this paper was obtained in the context of the I-CARE study [19]. I-CARE is a tablet-based activation system that provides user-specific activations such as memory games, rhymes, proverbs, image galleries or videos. The system is designed to be jointly used by tandems of PwD and formal/informal caregivers. Each participant with dementia fulfills the clinical criteria for dementia according to the ICD-10 system (Alzheimer dementia, vascular dementia, frontotemporal dementia, Korsakoff's syndrome, or Dementia Not Otherwise Specified). For the study, a setup with minimal supervision and setup requirement was selected with activation sessions taking place in private rooms or in commonly used spaces in care facilities. Audio recordings thus can contain non-stationary noise, e.g. (multiple) background speakers, audio-based activation contents or room reverberation, whilst video recordings may be impaired by lighting conditions or (partially) covered faces.

The dataset comprises 130 activation sessions with audio recordings, 113 (47:43 h) of which contain speech from the par-

http://dx.doi.org/10.21437/Interspeech.2021-567

ticipant with dementia. These sessions cover 24 PwD (gender: 15 f, 9 m; age: 58-94 years, *M*: 82.63 years, *SD*: 8.41 years; dementia stage: 8 mild-moderate, 5 severe, 11 unspecified). The audio (16 kHz) and video (30 FPS) signals were recorded using the tablet (Google Pixel C or Huawei MediaPad M5) microphone and camera, respectively. We manually annotate the dataset using the "Video Coding - Incorporating Observed Emotion" (VC-IOE) protocol [20]. Accordingly, emotional engagement is conceptualized based on five dimensions of affect (pleasure, anger, anxiety/fear, sadness or neutral). In the I-CARE context, however, certain negative responses (anger or fear) were not expected. We therefore aggregate the three negative states into one class covering overall negative affect. Verbal engagement is coded as present when the participant with dementia is participating in or maintaining a conversation. Annotation is performed retrospectively by two independent raters based on auditory or visual cues for each video-frame. We compute Cohen's Kappa (κ) between both raters after intensive training on six random test sessions to evaluate inter-rater reliability, and observe a high agreement of κ=0.824 for emotional and a substantial agreement of κ=0.783 for verbal engagement [21]. Tab. 1 summarizes the data that contains verbal (PwD speaks) and emotional engagement annotations (neutral, positive, negative) with regards to participants, sessions, utterances and video frames (Mio.). Most participants contribute multiple sessions (1-8 sessions, M:4.7) while each session contains multiple utterances which can be assigned to multiple video frames. Thus, individual utterances can contain multiple classes (see Fig. 2). In total, our dataset covers 24 participants who took part in 113 sessions which contain 1,051,654 frames with speech. Neutral (90.6 %) and positive (8.7 %) frames strongly outweigh negative (0.7 %) ones.

Table 1: *Class distribution with regard to participants, sessions, utterances and video frames (Mio.) expressed as frequencies and proportions (%).*

Label	*Part. (%)*	*Sess. (%)*	*Utt. (%)*	*(Mio.)Frames (%)*
Total	24 (100)	113 (100)	20,514 (100)	1.052 (100)
Neu.	24 (100)	113 (100)	18,796 (91.6)	0.953 (90.6)
Pos.	24 (100)	99 (87.6)	3,018 (14.7)	0.092 (8.7)
Neg.	9 (37.5)	24 (21.2)	142 (0.7)	0.007 (0.7)

Fig. 1 shows the distributions of frames that contain emotional (either positive or negative), verbal and emotional+verbal responses with regards to the stage of the disease. The results indicate that throughout all stages of the disease, participants show verbal and emotional responses to a similar degree. Emotional responses (auditory or visual) could be observed in 8.3 % of the frames (*SD*: 7.1 %) for unspecified, 9.3 % (*SD*: 7.7 %) for mild-moderate and 8.2 % (*SD*: 9.3 %) for severe dementia for all frames. Verbal responses were shown in 23.5 % (*SD*: 13.0 %) for unspecified, 29.7 % (*SD*: 13.6 %) for mild-moderate and in 21.3 % (*SD*: 19.0 %) of all frames for severe dementia. Emotional responses through or during speech could be found in 2.3 % (*SD*: 2.5 %) for unspecified, 2.3 % (*SD*: 1.2 %) for mild-moderate and 2.9 % (*SD*: 4.0 %) for severe dementia for all frames. Applying one-way ANOVA, statistically significant differences can be found between emotional responses (F=19.063, p<0.001) and stage, as same as verbal responses and stage (F=57.270, p<0.001). However, no statistically significant differences can be found between emotional responses through or during speech and stage (F==3.193, p=0.080). This indicates that regardless of the stage of the disease, PwD in our study were able to verbally express emotional responses, and

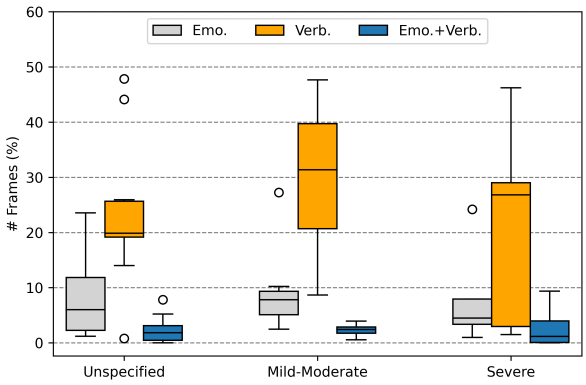

Figure 1: *Box-and-whisker plot of frames that contain emotional (positive, negative), verbal and emotional+verbal responses with regards to the stage of the disease.*

that speech remains a potential source of information for the recognition of emotional engagement. This result is in line with previous findings [22, 20] that facial expressions can become less or blunted throughout the progression of the disease, but stay intact for some PwD. At the same time, social interaction might further positively influence the expression [23].

4. Features and Methods

4.1. Pre-Processing

As data was collected in an unconstrained care setting (see Sec. 3), we apply denoising based on an encoder-decoder architecture [24] on all audio files to increase Signal-to-Noise Ratio (SNR). To obtain individual utterances of the participant with dementia, we first use a voice activity detection (VAD) on all sessions (Recall M: .76 SD: .08) based on Hidden-Markov-Models [25] which segment each session into speech and non-speech parts. Subsequently, speech segments (utterances) contributed by participants with dementia are manually assigned based on annotation data (see Sec. 3).

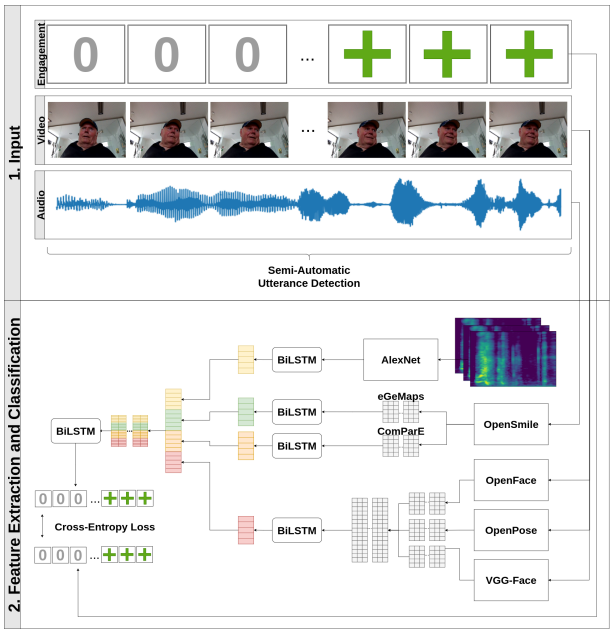

Figure 2: *Overview of our proposed system architecture.*

4.2. Feature Extraction

4.2.1. Visual Features

The face is arguably the most important non-verbal source for affective information [26]. Studies have also shown that facial expressions can be a promising indicator for affect of PwD despite age and disease-related changes [18]. Accordingly, we use the video signal to extract different feature sets to build a baseline system. Firstly, we detect, align and crop faces using OpenFace 2.0 [27]. Frames without a visible face are excluded from further analyses. Next, we extract facial features, namely the intensity of 17 Action Units (AUs) [1], the location and rotation of the head (head pose) and the direction of eye gaze for all video frames. We also use the pre-trained VGG-Face network to extract CNN features from each frame as proposed in [18]. Lastly, we extract skeleton features using OpenPose [28] to calculate relevant features, namely the distance between shoulders, eyes, ears, hands to nose and the visibility of the hands. Concatenating all mentioned features results in a 4137-d feature vector.

4.2.2. Audio Features

For each utterance, we extract three feature sets: (1) the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [7], (2) the 2013 Interspeech Computational Paralinguistics Challenge features set (ComParE) [8] and (3) the DeepSpectrum features [9, 29]. We use the OpenSMILE toolkit [30] to derive audio frame-wise frequency, energy and spectral related Low-Level Descriptors (LLD) for (1) and (2). This results in a 23-dimensional and a 130-dimensional feature vector per audio frame, respectively. For (3), we create mel spectrograms using Hanning windows with a size of 512 samples and an overlap of 256 samples. We forward spectrograms (227x227 pixels, viridis colormap) to the pre-trained AlexNet to receive bottleneck features from the fully connected layer (*fc7*) in accordance with [9]. This leads to a 4096-dimensional feature vector. Previous studies suggest that the minimum utterance length to reliably identify emotions from speech is about 250 ms [31, 32]. Thus, utterances with a shorter duration are excluded from classification. We apply z-normalization to audio and video features to consider the data variability that stems from the heterogeneity of the dataset.

4.3. Classification and Evaluation

We formulate our classification problem in a sequence-to-sequence manner, where the input is a sequence (250 ms with 50 % overlap during training) of feature vectors, which is mapped to a sequence of the labels for emotional engagement, at the rate of the video sampling rate, on which the ground truth labels are based. This architecture is based on Bidirectional Long-Short-Term-Memory Networks (BiLSTMs), which allow for the preservation of temporal dependencies. Concatenated visual features are fed into the visual encoder, which outputs a fixed dimensional context vector as a representation. Audio features are submitted to independent audio encoders which each output one context vector. Replacing unconsidered feature sets with vectors of zeros allows us to "mute" individual audio encoders, and thus to learn about their usefulness for the proposed task. Next, all context vectors are concatenated and passed over to the decoder, which outputs the label vector. We train the model for 50 epochs with a batch size of 6. We use cross-entropy loss function and Adam optimizer with a learning rate of 0.001. To smooth the model output, we use a median-filter of 15 frames. In total, we train six different models: (1) video-only as baseline (*Video*), (2-4) video and individual audio feature sets (see Sec. 4.2.2, (*V+eGeMAPS, V+ComParE, V+DeepSpectrum*)), (5) video and all audio feature sets combined (*V+Combined*) and (6) all audio feature sets combined without video (*Audio*). To find the best multimodal model, we individually selected the best model based on minimum training loss (*V+Best*). We perform user-dependent evaluations with a Leave-One-Session-Out (LOSO) cross-validation[2] due to the high data heterogeneity (recording settings, co-morbidities, stage and type of disease, medication) for all models. As only some participants show negative responses (see Tab. 1), we independently evaluate the two-class (neutral, positive) and three-class (neutral, positive, negative) problem. The latter is a subset of 5 PwD who show negative responses in at least two sessions. We select Unweighted Average Recall (UAR) as the evaluation metric as it is particularly suitable for imbalanced class distributions [9]. We apply pairwise t-test with Benjamini-Hochberg adjustment to test for statistical significance. For comparison of multiple groups, we apply one-way ANOVA.

5. Results and Discussion

Tab. 2 shows the classification results for the two-class problem as the *M* and *SD* UAR and recall score for each target class for all participants. The performance of the unimodal models (*Video: M*:.52, *SD*:.15; *Audio: M*:.52, *SD*:.03) is lower compared to all multimodal models. Best results are achieved based on *Video+Combined* (*M*:.57, *SD*:.08) and *Video+Best* (*M*:.57, *SD*:.09). Both models perform significantly better than the *Video* model (T=-3.481, $p<.05$; T=3.298, $p<.05$) and chance level (T=4.193, $p<.01$; T=4.071, $p<.01$). It becomes apparent that positive responses are best recognized by the *Video* model (*M*:.43, *SD*:.27) while recognition of neutral class works best for *Video+Combined* and *Video+Best* (*M*:.84, *SD*:.15). Tab. 3 provides the results for the three-class problem. The highest UAR is based on *Video+Combined* and *Video+Best* with the same scores (*M*:.41, *SD*:.04). Highest recall scores for the recognition of negative responses are reached based on *Video* (*M*:.11, *SD*:.10) and *Video+DeepSpectrum* (*M*:.11, *SD*:.12). Overall, these results show that audio features can help to improve recognition performance compared to a *Video* model. Accordingly, audio signal can be an important modality for the automatic recognition of emotional engagement of PwD. The best performance is achieved when using the combination of conventional acoustic feature sets with CNN-based features, which are assumed to be more robust to environmental noise [9]. Further, this combination helps to reduce the number of false positives which can be useful for an engagement-aware recommendation system. Activations that are perceived less positively than indicated by the system can interrupt the activation flow. A possible explanation for the low performance for negative responses is the little amount of training data for that class (see Tab. 1). It can further be assumed that participants showed only rather subtle negative emotional expressions due to the highly supportive social context. Thus, we expect that the average responses were more positive than if participants had engaged with the system entirely on their own [33]. We believe that the detection of engaging activations remains an

[1] AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45

[2] One participant was excluded because they only contributed a single session.

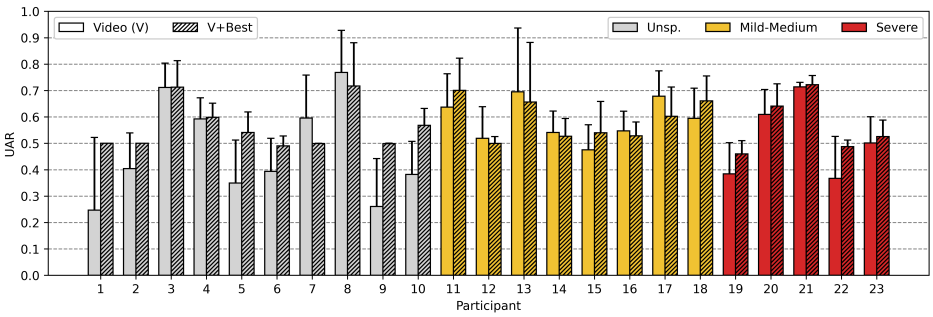

Figure 3: *Emotional engagement recognition results based on a LOSO cross-validation for the two-class problem. Results are reported as the M and SD UAR over all sessions of each participant. Plain bars represent video features while striped ones are based on video features and the best audio set. Bar colors represent the dementia stage (grey: unspecified, yellow: mild-medium, red: severe).*

Table 2: *Recognition results for the two-class problem based on all participants. Results for the different feature sets are reported as the M and SD recall score for each class over participants. Also, the UAR is reported.*

FeatureSet	*Neu.*	*Pos.*	*UAR*
Audio	.79 (.23)	.24 (.25)	.52 (.03)
Video (V)	.61 (.21)	**.43 (.27)**	.52 (.15)
V+eGeMAPS	.74 (.18)	.38 (.27)	.56 (.10)
V+ComParE	.79 (.17)	.34 (.27)	.56 (.09)
V+DeepSpec.	.74 (.20)	.37 (.28)	.55 (.10)
V+Combined	**.84 (.15)**	.31 (.27)	**.57 (.08)**
V+Best	**.84 (.15)**	.31 (.27)	**.57 (.09)**
Chance	1.0 (.00)	.00 (.00)	.50 (.00)

Table 3: *Recognition results for the three-class problem based on 5 PwD that show negative responses in at least two sessions.*

FeatureSet	*Neu.*	*Pos.*	*Neg.*	*UAR*
Audio	.57 (.29)	.46 (.29)	.01 (.02)	.34 (.01)
Video (V)	.46 (.23)	**.51 (.12)**	**.11 (.10)**	.36 (.09)
V+eGeMAPS	.60 (.16)	.46 (.16)	.08 (.09)	.38 (.07)
V+ComParE	.63 (.21)	.47 (.27)	.07 (.15)	.39 (.05)
V+DeepSpec.	.58 (.16)	.48 (.21)	**.11 (.12)**	.39 (.08)
V+Combined	.75 (.09)	.45 (.20)	.04 (.05)	**.41 (.04)**
V+Best	**.76 (.10)**	.44 (.18)	.04 (.05)	**.41 (.04)**
Chance	1.0 (.00)	.00 (.00)	.00 (.00)	.33 (.00)

important challenge. This is why we aim to provide this functionality for a real-world, and thus noisy and unconstrained care setting, and this scenario is simply not comparable to the type of clean and unambiguous data that can be obtained in (laboratory) studies with healthy individuals [11]. Finally, PwD undergo substantial age-related changes, e.g. development of folds and wrinkles, whilst also often suffering from prosody impairment and reduced affective expressiveness [6, 20]. Our present results therefore take an important step towards achieving a robust, real-world classification of emotional engagement for PwD.

Fig. 3 shows the user-dependent recognition performance for *Video* and *Video+Best* with regard to the dementia stage (grey: unspecified, yellow: mild-medium, red: severe). When we compare the means of these groups, we can find statistically significant differences for *Video* ($F=9.701$, $p<.01$) and *Video+Best* ($F=8.59$, $p<.01$). Interestingly, group means especially increase for unspecified (M:.47, SD:.18) when adding audio signal (M:.56, SD:.09) and for severe dementia (M:.52, SD:.15 to M:.57, SD:.11). It also becomes apparent that classification performance varies substantially between participants. These differences appear likely to be the result of on interplay of several factors, including personality, recording conditions or the specific type of dementia for example, rather than the gross severity of the disease as such.

6. Conclusions

The main aim of this study was to examine the usefulness of speech for the automatic recognition of engagement for PwD. Our results show that speech can significantly improve automatic classification based on video signals. The collection of speech data is non-intrusive and can thus make an important contribution towards detecting engagement of the PwD. Intriguingly, we were able to obtain promising recognition results even for some participants in a severe stage of dementia. Consistent with prior findings, e.g. [22], this suggests that at least some PwD remain sufficiently expressive for our multimodal approach to accurately infer their emotional engagement despite all impediments. As shown by the significant performance improvements due to the addition of speech, we argue that models based on multiple modalities are likely to be most promising with respect to future advances in this field. In further research, we aim to also investigate other dimensions of engagement [20] to obtain a more comprehensive picture of the behavioral dynamics taking place in the activation of PwD.

7. Acknowledgements

This work was partially funded by the Klaus-Tschira-Stiftung. Data collection and development of the I-CARE system was funded by the BMBF under reference BMBF-number V4PIDO62. We also gratefully acknowledge the support of the Leibniz ScienceCampus Bremen Digital Public Health (lsc-diph.de), which is jointly funded by the Leibniz Association (W4/2018), the Federal State of Bremen and the Leibniz Institute for Prevention Research and Epidemiology – BIPS.

8. References

[1] WHO, "Dementia," https://www.who.int/news-room/fact-sheets/detail/dementia, 2017, [Online; accessed 08-October-2020].

[2] B. Woods, E. Aguirre, A. E. Spector, and M. Orrell, "Cognitive stimulation to improve cognitive functioning in people with dementia," *Cochrane Database of Systematic Reviews*, no. 2, 2012.

[3] A. Spector, L. Thorgrimsen, B. Woods, L. Royan, S. Davies, M. Butterworth, and M. Orrell, "Efficacy of an evidence-based

cognitive stimulation therapy programme for people with dementia: randomised controlled trial," *The British Journal of Psychiatry*, vol. 183, no. 3, pp. 248–254, 2003.

[4] J. Cohen-Mansfield, M. Dakheel-Ali, and M. S. Marx, "Engagement in persons with dementia: the concept and its measurement," *The American journal of geriatric psychiatry*, vol. 17, no. 4, pp. 299–307, 2009.

[5] K. Horley, A. Reid, and D. Burnham, "Emotional prosody perception and production in dementia of the alzheimer’s type," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 5, pp. 1132–1146, 2010.

[6] V. J. Roberts, S. M. Ingram, M. Lamar, and R. C. Green, "Prosody impairment and associated affective and behavioral disturbances in alzheimer's disease," *Neurology*, vol. 47, no. 6, pp. 1482–1488, 1996. [Online]. Available: https://n.neurology.org/content/47/6/1482

[7] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April 2016.

[8] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *MM '13*, 2013.

[9] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 478–484. [Online]. Available: https://doi.org/10.1145/3123266.3123371

[10] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56 – 76, 2020.

[11] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, p. 90–99, Apr. 2018. [Online]. Available: https://doi.org/10.1145/3129340

[12] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 674–681.

[13] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017 - real-life depression, and affect recognition workshop and challenge," in *AVEC'17*. United States: Association for Computing Machinery (ACM), 10 2017, pp. 3–9.

[14] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller, "Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 1–5.

[15] B. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing masks," 2020.

[16] D. S. Nazareth, "Emotion recognition in dementia: Advancing technology for multimodal analysis of emotion expression in everyday life," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019, pp. 45–49.

[17] K. Ma, X. Wang, X. Yang, M. Zhang, J. M. Girard, and L.-P. Morency, "Elderreact: A multimodal dataset for recognizing emotional response in aging adults," in *2019 International Conference on Multimodal Interaction*, ser. ICMI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 349–357. [Online]. Available: https://doi.org/10.1145/3340555.3353747

[18] L. Steinert, F. Putze, D. Küster, and T. Schultz, "Towards engagement recognition of people with dementia in care settings," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 558–565.

[19] T. Schultz, F. Putze, T. Schulze, L. Steinert, R. Mikut, W. Doneit, A. Kruse, A. Depner, I. Franz, M. Engels, P. Gaerte, S. Jünger, R. Linden, C. Ziegler, M. Ricken, T. Dimitrov, J. Herzig, I. Maucher, K. Bernardin, and C. Simon, "I-care - ein menschtechnik interaktionssystem zur individuellen aktivierung von menschen mit demenz," 06 2018.

[20] C. Jones, B. Sung, and W. Moyle, "Assessing engagement in people with dementia: a new approach to assessment using video analysis." *Archives of psychiatric nursing*, vol. 29 6, pp. 377–82, 2015.

[21] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Family medicine*, vol. 37, no. 5, pp. 360–363, May 2005. [Online]. Available: http://europepmc.org/abstract/MED/15883903

[22] C. Magai, C. Cohen, D. Gomberg, C. Malatesta, and C. Culver, "Emotional expression during mid- to late-stage dementia," *International Psychogeriatrics*, vol. 8, no. 3, p. 383–395, 1996.

[23] K. H. Lee, M. Boltz, H. Lee, and D. Algase, "Does social interaction matter psychological well-being in persons with dementia?" *American Journal of Alzheimer s Disease and Other Dementias*, vol. 32, p. 153331751770430, 04 2017.

[24] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[25] D. Telaar, M. Wand, D. Gehrig, F. Putze, C. Amma, D. Heger, N. T. Vu, M. Erhardt, T. Schlippe, M. Janke, C. Herff, and T. Schultz, "Biokit - real-time decoder for biosignal processing," in *The 15th Annual Conference of the International Speech Communication Association, Singapore*, 2014, interspeech 2014.

[26] A. Kappas, E. Krumhuber, and D. Küster, "Facial behavior," in *In: Hall, Judith A.; Knapp, Mark L. (Ed.), Nonverbal communication (S. 131-166). Berlin: de Gruyter, 2013.* de Gruyter, 2013, pp. 131–166.

[27] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[28] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[29] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. Interspeech 2017*, 2017, pp. 3512–3516. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-434

[30] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[31] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 3682–3686.

[32] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153 – 163, 2013, affect Analysis In Continuous Input.

[33] A. Fridlund, "Sociality of solitary smiling: Potentiation by an implicit audience," *Journal of Personality and Social Psychology*, vol. 60, pp. 229–240, 02 1991.