Speech Emotion Recognition among Elderly Individuals using Multimodal Fusion and Transfer Learning

George Boateng
gboateng@ethz.ch
ETH Zürich
Zurich, Switzerland

Tobias Kowatsch
tkowatsch@ethz.ch
ETH Zürich
Zurich, Switzerland
University of St. Gallen
St. Gallen, Switzerland

ABSTRACT

Recognizing the emotions of the elderly is important as it could give an insight into their mental health. Emotion recognition systems that work well on the elderly could be used to assess their emotions in places such as nursing homes and could inform the development of various activities and interventions to improve their mental health. However, several emotion recognition systems are developed using data from younger adults. In this work, we train machine learning models to recognize the emotions of elderly individuals via performing a 3-class classification of valence and arousal as part of the INTERSPEECH 2020 Computational Paralinguistics Challenge (COMPARE). We used speech data from 87 participants who gave spontaneous personal narratives. We leveraged a transfer learning approach in which we used pretrained CNN and BERT models to extract acoustic and linguistic features respectively and fed them into separate machine learning models. Also, we fused these two modalities in a multimodal approach. Our best model used a linguistic approach and outperformed the official competition of unweighted average recall (UAR) baseline for valence by 8.8% and the mean of valence and arousal by 3.2%. We also showed that feature engineering is not necessary as transfer learning without fine-tuning performs as well or better and could be leveraged for the task of recognizing the emotions of elderly individuals. This work is a step towards better recognition of the emotions of the elderly which could eventually inform the development of interventions to manage their mental health.

CCS CONCEPTS

• **Applied computing** → **Psychology**.

KEYWORDS

Speech emotion recognition; Affective computing; Transfer learning; Computational paralinguistics; Elderly individuals; Multimodal fusion; Deep learning; CNN; LSTM; BERT; SBERT; Support vector machine

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425255

ACM Reference Format:
George Boateng and Tobias Kowatsch. 2020. Speech Emotion Recognition among Elderly Individuals using Multimodal Fusion and Transfer Learning. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425255

1 INTRODUCTION

Digital technologies are needed to aid in managing the physical and emotional well-being of elderly individuals [24]. Awareness of the emotions of the elderly could give an insight into their mental health. Emotion recognition systems that work well on the elderly could be used to assess their emotions in places such as nursing homes and could inform the development of various activities and interventions to improve their mental health. However, several emotion recognition works use data collected from actors and younger adults for their development and evaluation (e.g. IEMOCAP dataset [3]). In this work, we develop and evaluate emotion recognition models using the first public speech data collected from elderly individuals in the real world for emotion recognition as part of the INTERSPEECH 2020 Computational Paralinguistics Challenge (COMPARE) [26]. The task was to perform a 3-class classification of the arousal and valence dimensions of emotions based on speech data from elderly individuals.

Deep learning has been used for speech emotion recognition involving various approaches such as convolutional neural networks (CNN), Recurrent Neural Networks (RNN) such as Long Short-Term Memory (LSTM) — with and without attention — bidirectional LSTM (BLSTM), mostly together with handcrafted features ([17]). Other approaches have used the raw signal in an end-to-end approach leveraging 1D CNNs and LSTMs [29]. Transfer learning is another approach used in deep learning to circumvent the need to develop hand-crafted features and also deals with the challenge off small labeled datasets. Transfer learning entails pretraining a model on a different but related task and using it for feature extraction or fine-tuning in which the whole model or later layers are retrained ([7]). Transfer learning has shown success in various fields such as computer vision ([13, 16]), speech processing ([15]), and natural language processing ([12, 23]). Transfer learning has also been used in emotion recognition tasks ([7, 14, 25]).

Our contribution is the evaluation of transfer learning approaches to recognize the emotions of elderly individuals using a novel dataset — speech data collected from German-speaking elderly individuals. Specifically, we used a pretrained CNN model to extract

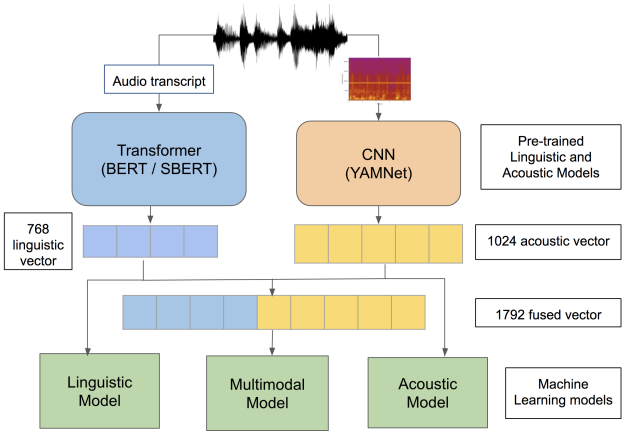

Figure 1: Overview of Acoustic, Linguistic and Multimodal Approaches

acoustic features and a pretrained Transformer language model — Bidirectional Encoder Representations from Transformers (BERT) [6] — to extract linguistic features. We trained and evaluated separate models for acoustic and linguistic modalities. Also, we used a multimodal approach in which we fused the features (early fusion) and trained models using the combined features [20].

The rest of our paper is organized as follows. In Section 2, we describe our methodology. In Section 3, we describe our experiments. In Section 4, we show the results, discuss them and present future work. We conclude in Section 5.

2 METHODS

In this section, we describe the dataset, the competition baseline approaches, and our acoustic, linguistic, and multimodal approaches as shown in Figure 1.

2.1 Dataset

We used the Ulm State of mind elderly (USOMS-e) database collected from German-speaking elderly individuals [26]. The dataset contains speech data of 87 participants (55 f, 32 m, age 60–95 years, mean 71.01 years, std. dev. 9.14 years), each of whom told two negative and one positive personal narrative. Participants' emotions were assessed post every narrative by the subject and later by 4 experts on a scale of 0 (very sleepy and very bad) to 10 (very excited and very good) for the "arousal" and "valence" dimensions respectively. The audio data was converted to 16 KHz mono and was segmented into 5-sec chunks. The audio was also transcribed manually and automatically. The mean values of each dimension were used to create 3 classes: low (0-6), medium (7-8), and high (9-10).

2.2 Competition Baseline Approach

The organizers of the competition used various approaches to generate the baseline results for the competition such as feature engineering, transfer learning, unsupervised learning and end-to-end learning [26]. For feature engineering, they used the openSMILE toolkit to extract 6373 static features (functionals), and the OPENXBOW toolkit to extract Bag of Audio Words (BoAW) features. For transfer learning, they used the DEEP SPECTRUM toolkit which used a pretrained CNN (ResNet50) to extract embeddings from the spectrograms of the audio. They also used the LinguistIc Feature Extractor (LIFE) toolkit to extract linguistic embeddings which used a BERT model that was pretrained on German text followed by Global Maximum pooling or bidirectional LSTM with attention. For unsupervised learning, they used the AUDEEP toolkit which used recurrent sequence-to-sequence autoencoders to learn representations of mel-spectrograms of the audio in an unsupervised manner. These different feature sets were then fed into separate linear support vector machines with different hyperparameters.

2.3 Acoustic Approach

We used the acoustic characteristics of the audio to perform classification. We extracted spectrograms and used a pretrained CNN to compute embeddings which we used as acoustic features to perform classification with various machine learning models (Figure 1). We used the YAMNet model which is a CNN that was pretrained on the AudioSet dataset to predict 521 audio event classes [8, 9]. YAMNet is based on the MobileNet architecture [11]. We used the YAMNet model as a feature extractor and hence replaced the original final logistic layer which outputs 521 class with various machine learning algorithms.

We extracted a spectrogram as an input into the YAMNet model in the same way as was done for the trained model. The audio is downsampled from 44.1Kz to 16 kHz mono. A spectrogram is computed using magnitudes of the Short-Time Fourier Transform with a window size of 25 ms, a window hop of 10 ms, and a periodic Hann window. A mel spectrogram is computed by mapping the spectrogram to 64 mel bins covering the range 125-7500 Hz. A stabilized log mel spectrogram is computed by applying log(mel-spectrum + 0.01) where the offset is used to avoid taking a logarithm of zero. These features are then framed into non-overlapping examples of 0.96 seconds, where each example covers 64 mel bands and 96 frames of 10 ms each. This resulted in a 2D data of size 96 x 64 for each second, which we used as a data point input to the YAMNet model. The output of the model is a 1024-dimensional feature vector per data point input of size 96 x 64. We then normalized the feature vectors to be zero mean and unit variance and then used them as inputs to various machine learning models.

2.4 Linguistic Approach

We used the content of the speech — the manual transcript — to perform classification. Specifically, we used pretrained Transformer language models to extract linguistic features and then performed classification with various models (Figure 1). We used a pretrained BERT model to extract a 768-dimensional embedding vector for each narrative [6]. BERT is a deep learning model that has achieved state-of-the-art results for several natural language tasks. The BERT model we used is a case sensitive German BERT that was trained using a German Wikipedia dump, the OpenLegalData dump, and news articles [2]. We preprocessed each story's transcript by first tokenizing each word and ensuring that the total number of tokens was less than or equal to the 512 maximum that the BERT model takes. Hence, we ignored subsequent words in each story which was

Table 1: Results for Competition Baseline Approaches and our Acoustic, Linguistic, and Multimodal Approaches

Model	Dev (UAR %)		Test (UAR %)	
	Val	Arous	Val	Arous
Competition Baseline Approach				
Functionals + SVM	33.3	39.1	33.3	47.9
BoAW + SVM	33.3	40.5	31.5	49.1
Autoencoder + SVM	36.7	34.9	33.8	44.3
ResNet50 + SVM	31.6	35.0	40.3	**50.4**
BERT + LSTM + SVM	49.2	40.6	**49.0**	44.0
Acoustic Approach				
YAMNet + SVM	44.3	43.9	34.7	43.9
YAMNet + LSTM	37	40.2	—	47.9
Linguistic Approach				
BERT + SVM	51.1	45.7	56.3	**48**
SBERT + SVM	57.42	30.33	**57.8**	—
Multimodal Approach				
Fusion + SVM	49	43.8	52.3	47.4

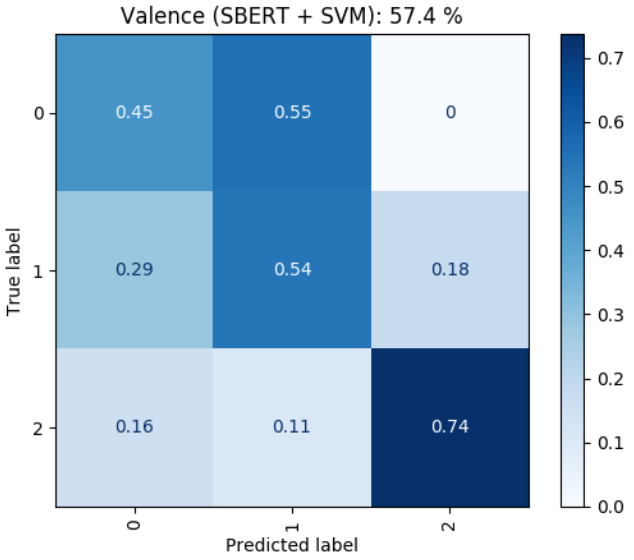

Figure 2: Confusion matrix for development set evaluation for the best model for valence — SBERT + SVM: 57.4%

over 512 length. We added special tokens for sentence classification (such as [CLS] at the first position). After passing each story into the model, we took the 768-dimensional embedding vector of the first token [CLS] of the last hidden layer and used that as the embedding for the whole story. We then normalized the vectors to be zero mean and unit variance and then used the features vectors as inputs to various machine learning models.

We also used Sentence BERT (SBERT), a modification of the BERT architecture with siamese and triplet network structures for generating sentence embeddings such that semantically similar sentences are close in vector space [21]. The SBERT network was shown to outperform state-of-the-art sentence embedding methods such as BERT and Universal Sentence Encoder for semantic similarity and sentence classification tasks such as sentiment detection. We used the multilingual version of the SBERT model [22]. The network, like the original BERT outputs a 768-dimensional embedding for each story. We normalized the vectors to be zero mean and unit variance and then used the feature vectors as inputs to various machine learning models.

2.5 Multimodal Approach

We also explored using a multimodal approach in which we fused aspects of the acoustic and linguistic modalities because multimodal approaches have been shown to outperform unimodal approaches in emotion recognition tasks [20]. Specifically, we fused the feature vectors from the acoustic and linguistic approaches producing a 1792-dimensional feature vector for each story (Figure 1). Since there were several acoustic feature vectors for each story, we performed a weighted sum of the acoustic feature vectors for each story. We then normalized the vectors to be zero mean and unit variance and then used these fused vectors as inputs to various machine learning models.

3 EXPERIMENTS

We performed various experiments using the following libraries scikit-learn [19], keras [5], and PyTorch [18]. We trained models separately for valence and arousal, and used a hyperparameter search to get models that produced the best results. We used a linear support vector machine (SVM), and a 2-layer LSTM [10] with 16 and 8 units, and 50% dropout [27] after each layer. We used the LSTM model for the acoustic approach to take advantage of the sequential nature of the acoustic embeddings. Also, for the acoustic approach, we used majority voting of the classification of the 5-sec audio chunks to decide the class for each story. For evaluation, we used the metric unweighted average recall (UAR) which is used for unbalanced data and confusion matrices.

Given that the data is imbalanced, we upsampled the minority classes so the data was balanced using the SMOTE algorithm [28] and imblearn library [1]. We used the train and development data sets provided by the competition organizers for developing the model. The organizers had a held-out test whose labels were not made available to researchers. We had to submit our predictions on the held-out test which was evaluated by the organizers, and the prediction result sent to us. Also, we had a constraint of five submissions on the held-out test set and hence we used only our best models for those submissions. The official competition baseline was based on the performance on the held-out test set.

4 RESULTS, DISCUSSION AND FUTURE WORK

We present the results for the competition baseline, and our acoustic, linguistic, and multimodal approaches in Table 1 where a "—" means that the model was not used for the held-out test. The best results for the competition baseline and our approaches in the valence and arousal columns are highlighted in **bold**. Also, we show the

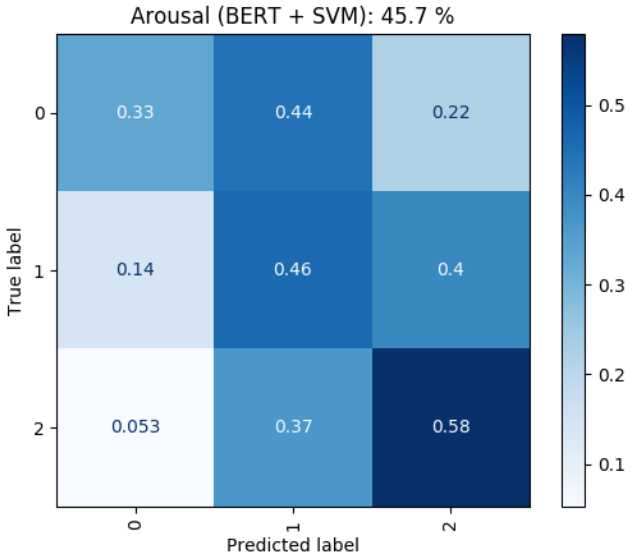

Figure 3: Confusion matrix for development set evaluation for the best model for arousal: BERT + SVM: 45.7%

confusion matrices of the best models in Figure 2 (valence) and 3 (arousal).

The competition organizers' best methods which produced the results that were used as the official competition baseline results were the DEEP SPECTRUM (ResNet50 + SVM) for acoustic (50.4%) and LIFE (BERT + LSTM + SVM) for valence (49.0%) and an average of valence and arousal of 49.7%.

Among our approaches, the linguistic models performed the best for both valence and arousal, with the multimodal model being the second best for valence and the acoustic model being the second best for arousal. Our best model for valence was SBERT + SVM with a UAR of 57.8% and the best model for arousal is BERT + SVM with UAR of 48% and an overall mean UAR of valence and arousal being 52.9%. Our best models outperformed the official baseline (using the held-out test set) for valence by 8.8% and the mean of valence and arousal by 3.2%. Our best arousal model is however below the official arousal baseline by 2.4%. Our acoustic models not performing better than the baseline suggests that using the pretrained YAMNet model as feature extractor is not adequate. Hence, fine-tuning the model additionally or pretraining the model on a related emotion recognition task might be necessary for good performance.

The linguistic model performing better than the acoustic model is consistent with the results of other works such as an emotion recognition task among real-world couples whose best recognition result for a 3-class classification of valence was 57.42% (UAR) [4]. A possible explanation is that we used the manual transcript which is a perfect representation of the narratives which the linguistic model used as compared to the acoustic models which worked on raw, noisy, audio data. The model might have performed poorly with the automatic transcript but we did not evaluate that as we used only the best model for evaluation. Also, the SBERT model performed better than the regular BERT model for valence. This result is consistent with [21] which showed that SBERT extracts better sentence embeddings than BERT for sentiment detection tasks.

The multimodal model surprising did not perform the best considering multimodal approaches have been shown to perform better than unimodal approaches. This performance is however consistent with the result of [4]. It is possible that the limitations of the acoustic features affected the multimodal results since we performed feature-level fusion. Exploring other forms of fusion like decision-level and hybrid may improve the results of the multimodal approach.

Our transfer learning approaches performed as well or better than the competition baseline approaches that used feature engineering (static features and BoAW). These results show that feature engineering is not necessary to get good emotion classification results for real-world speech data from older adults. This work focused on using pretrained models as feature extractors. Hence, we did not fine-tune the pretrained YAMNet and BERT models on this data. Doing so in the future could improve the recognition results.

Finally, this work is a key step towards recognizing the emotions of elderly individuals in daily life. We have collected speech and video data with self-reported emotion labels from German-speaking elderly individuals in their daily life after they underwent inpatient cardiovascular rehabilitation. Our future work will build upon this work and explore emotion recognition within that unique context.

5 CONCLUSIONS

In this work, we used a transfer learning approach to classify low, medium, and high emotion labels of the valence and arousal dimension of audio data collected from German-speaking elderly individuals. We used pretrained CNN and BERT models to extract acoustic and linguistic features respectively and fed them into separate machine learning models. Additionally, we fused the features in a multimodal approach and fed them to machine learning models. Our models using a linguistic approach performed better than the official competition baseline for the valence recognition task by 8.8%. Also, our results showed that feature engineering is not necessary and transfer learning can be leveraged to produce decent performance for the task of recognizing the emotions of elderly individuals. This work is a step towards better recognition of the emotions of the elderly which could eventually inform the development of interventions to manage their mental health.

REFERENCES

[1] [n.d.]. imbalanced-learn. https://imbalanced-learn.readthedocs.io/en/stable/index.html. Accessed: 2020-05-1.
[2] [n.d.]. Open Sourcing German BERT. https://deepset.ai/german-bert. Accessed: 2020-05-1.
[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.
[4] Sandeep Nallan Chakravarthula, Haoqi Li, Shao-Yen Tseng, Maija Reblin, and Panayiotis Georgiou. 2019. Predicting Behavior in Cancer-Afflicted Patient and Spouse Interactions using Speech and Language. (2019).
[5] François Chollet et al. 2018. Keras: The python deep learning library. *Astrophysics Source Code Library* (2018).
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Kexin Feng and Theodora Chaspari. 2020. A Review of Generalizable Transfer Learning in Automatic Emotion Recognition. *Frontiers in Computer Science* 2 (2020), 9.

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[9] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[12] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[13] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[14] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 443–449.

[15] Charles C. Onu, Jonathan Lebensold, William L. Hamilton, and Doina Precup. 2019. Neural Transfer Learning for Cry-Based Diagnosis of Perinatal Asphyxia. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, Gernot Kubin and Zdravko Kacic (Eds.). ISCA, 3053–3057. https://doi.org/10.21437/Interspeech.2019-2340

[16] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1717–1724.

[17] Sandeep Kumar Pandey, HS Shekhawat, and SRM Prasanna. 2019. Deep Learning Techniques for Speech Emotion Recognition: A Review. In *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 1–6.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 8024–8035.

[19] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[20] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3973–3983.

[22] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *arXiv preprint arXiv:2004.09813* (04 2020). http://arxiv.org/abs/2004.09813

[23] Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpiDEDIS at GermEval 2019: Offensive Language Identification using a German BERT model. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). Erlangen, Germany: German Society for Computational Linguistics & Language Technology*. 403–408.

[24] Wendy A Rogers and Tracy L Mitzner. 2017. Envisioning the future for older adults: Autonomy, health, well-being, and social connectedness with technology support. *Futures* 87 (2017), 133–139.

[25] Sourav Sahoo, Puneet Kumar, Balasubramanian Raman, and Partha Pratim Roy. 2019. A Segment Level Approach to Speech Emotion Recognition Using Transfer Learning. In *Asian Conference on Pattern Recognition*. Springer, 435–448.

[26] Bjorn W. Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. In *Proceedings of Interspeech*. Shanghai, China, 5 pages. to appear.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[28] Juanjuan Wang, Mantao Xu, Hui Wang, and Jiwu Zhang. 2006. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In *2006 8th international Conference on Signal Processing*, Vol. 3. IEEE.

[29] Zixiaofan Yang and Julia Hirschberg. 2018. Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks.. In *Interspeech*. 3092–3096.