INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech-Based Classification of Defensive Communication: A Novel Dataset and Results

Shahin Amiriparian[1], *Lukas Christ*[1], *Regina Kushtanova*[1,2], *Maurice Gerczuk*[1], *Alexandra Teynor*[2], *Björn W. Schuller*[1,3]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Institute for Agile Software Development, Augsburg University of Applied Sciences, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

shahin.amiriparian@uni-a.de

Abstract

Defensive communication is known to have detrimental effects on the quality of social interactions. Hence, recognising and reducing defensive behaviour is crucial to improving professional and personal communication. We introduce DefComm-DB, a novel multimodal dataset comprising video recordings in which one of the following types of defensive communication is present: (i) verbally attacking the conversation partner, (ii) withdrawing from the communication, (iii) making oneself greater, and (iv) making oneself smaller. Subsequently, we present a machine learning approach for the automatic classification of DefComm-DB. In particular, we utilise wav2vec2, autoencoders, a pre-trained CNN and openSMILE for feature extraction from the audio modality. For the text stream, we apply ELECTRA and SBERT. On the unseen test set, our models achieve an Unweighted Average Recall of 49.4 % and 52.2 % for the audio and text modalities, respectively, showing the feasibility of the introduced challenge.
Index Terms: speech processing, defensive communication, computational paralinguistics, Transformers

1. Introduction

Effective communication plays an essential role in various aspects of human life, including personal and professional development, relationships, and overall well-being. Positive interpersonal communication helps to establish understanding and trust between individuals, making interaction more harmonious [1]. However, due to its complexity, interpersonal communication can fail in many ways. Contributing factors can include cultural and linguistic differences, societal norms, roles and values [2], stress level, mental health concerns, and personal circumstances [3]. In addition, various communication flaws can cause messages to be misinterpreted and expressed in an ineffective manner [4].

One particular form of behaviour in communication is defensiveness. It is often characterised by hostile, aggressive, or passive-aggressive behaviour [5] and can manifest as verbal and nonverbal responses [5] and actions directed toward reducing anxiety [6]. According to Stamp et al. [7], defensive behaviour comprises three parts: a self-perceived flaw that one refuses to acknowledge, sensitivity to that flaw, and a perceived or actual attack by another individual focused on that flaw [7]. This type of behaviour can reduce the effectiveness of communication [8] and cause defensive or aggressive responses in return which can lead to a detrimental cycle of poor communication [5]. Furthermore, defensiveness can negatively affect the quality and satisfaction of relationships and romantic partnerships by increasing the number of conflicts [9, 10]. Moreover, it has been identified as one of the key factors in marriage breakdown [9]. In the workplace, it can have a negative impact on the quality of leader-member interaction which is associated with higher levels of burnout and lower job satisfaction [11]. It is thus important to understand the dynamics of defensive communication and be able to quickly identify when it emerges and steer conversations constructively. By detecting defensive behaviour, individuals can navigate conflicts, resolve them, improve communication skills, and interact better in society. Therefore, it is desirable to have systems in place that can assist in this process.

So far, prior studies on this topic are non-computational and mainly questionnaire-based. Further, they have primarily focused on defence mechanisms, a phenomenon that underlies defensiveness. Various tools have been developed to study defensive behaviours that a person may not be aware of, based on Freud's idea that an observer can reliably determine them [12]. Such tools include the Lerner Defence Scale [13] based on the Rorschach inkblot test and a method developed by Cooper et al. [14] that measures 15 defences from three categories. Following the hierarchical organisation of defence mechanisms proposed by Vaillant et al. [15], Perry et al. [16] developed the Defense Mechanisms Rating Scale [16] and its computerised Q-sort version Q-sort based Defense Mechanisms Rating Scale (DMRS-Q) [17], which provide quantitative and qualitative scores reflecting a person's defensive functioning.

To the best of our knowledge, there are no approaches for automatic classification of defensiveness and defence mechanisms in communication. Other studies which are mainly from the field of computational paralinguistics [18, 19] explore each individual's character traits such as Big Five, insecurity, social anxiety, and feeling threatened that can contribute to defensive behaviour [7].

To this end, in a first attempt to automatic defence mechanism classification, we introduce a novel dataset for defensive communication (cf. Section 2) and provide machine learning solutions for its automatic classification (cf. Section 3). Applications of our research include but are not limited to the fields of conflict resolution [20], communication training [21], and psychology [22].

2. Dataset

We have collected a novel Multimodal **Def**ensive **Comm**unication Database, denoted as DEFCOMM-DB from YouTube. It comprises genuine *non-acted* dialogues between English-speaking individuals in 'real-world' settings that feature one of the defensive behaviours outlined in Birkenbihl's model of communication failures [21]:

1. Attacking the conversation partner (class *Attack*): videos that depict individuals actively attacking verbally, blaming the other person, or shifting the other person's attention to them-

10.21437/Interspeech.2023-76

selves.

2. Withdrawing from the communication (class *Flight*): videos where people refuse to respond, withdraw from the conversation or change the topic or focus.
3. Making oneself greater (class *Greater*): videos that depict individuals boasting, self-justifying in an aggressive manner, denying accusations, exhibiting a sense of dominance or superiority, or expressing indignation.
4. Making oneself smaller (class *Smaller*): videos that display individuals engaging in self-deprecation, self-blame, exhibiting a sense of guilt, apologising, and expressing feelings of vulnerability or worthlessness.

Initially, we collected 431 data points. Next, we carefully checked the collected videos to ensure that they aligned with Birkenbihl's model [21] and that no samples were repeated. After this evaluation, a final set of 261 data points was created. Key statistics on the dataset are provided in Table 1. DEFCOMM-DB features a variety of video topics, including interviews with celebrities and professional athletes, political debates, legal trials, TV shows, and video footage obtained by paparazzi, among others. The situations, number of participants, gender, age, and ethnicity vary from scene to scene. From each video, we retrieve audio, visual, and textual modalities. In this paper, we focus on the audio modality and the speech transcriptions. We have made DEFCOMM-DB available for academic researchers[1].

2.1. Search Parameters

We based our search for suitable YouTube videos on the assumption that defensive behaviour can be identified by certain verbal and non-verbal signs. Verbal indicators of defensive communication may include making excuses, denying responsibility [23], shifting blame, interrupting others [10], and using loud, fast, aggressive, or monotonous and evaluative speech [24]. Non-verbal cues of defensiveness may include closed body posture, avoidance of eye contact, specific head positioning, and facial expressions that convey hostility or disinterest [5]. To find these cues, we look for situations that could potentially involve conflict. Examples of search queries included but were not limited to: "bad interviews", "heated sports interviews", "celebrities freak out", "emotional court speeches", "heated debates", "bragging", "shaming", "refusal to answer", and "painful conversations".

2.2. Annotation

To build the gold standard, the video recordings were annotated by 11 participants (5 f and 6 m) with a mean age of 25 years ($\pm$ 4.9 years). Among them, four are AI researchers, one master's student with a background in psychology, and six are undergraduates studying social science and psychology (2), computer science (2), economics (1), and management (1). Prior to the annotation process, the participants were provided with verbal and written instructions outlining the purpose and goals of the study, a detailed description of the four classes that needed to be labelled, examples of videos featuring targeted reactions, and technical information about the labelling process. The annotators then assigned each data point to one of the following labels: *Attack*, *Smaller*, *Greater*, and *Flight*. Additionally, they could select if the video does not have any of the targeted reactions and leave a comment if they noticed some issues.

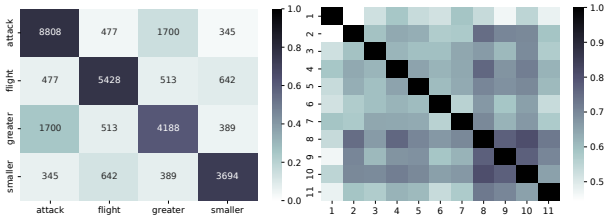
[1] https://zenodo.org/record/7706919

Table 1: *Statistics on* DEFCOMM-DB*: number of video clips, mean duration (μ), standard deviation (σ), minimum, maximum, and total duration of collected videos per class.*

Label	**# video clips**	μ [s]	σ [s]	**min** [s]	**max** [s]	Σ **duration** [s]
Attack	112	8	9	2	46	949
Flight	57	9	8	2	62	494
Greater	45	9	6	2	25	416
Smaller	47	12	8	3	49	556
Total	**261**	**9**	**8**	**2**	**62**	**2415**

2.2.1. Inter-Annotator Agreement

A large degree of subjectivity may be involved in the annotation process, even if it is based on clear instructions. Therefore, to ensure the reliability of the dataset, it is necessary to calculate the level of agreement between annotators. To do this, we use Krippendorf's alpha (α) [25]. For our collected annotations, we obtain an α value of 0.63, indicating the difficulty of the labelling and substantial confusion between annotators. As shown in Figure 1a, there is a more prominent tendency for confusion between the classes *Greater* and *Attack*, as well as between *Smaller* and *Flight*. Despite this confusion, it can be noted that overall, the annotators tend to label samples in a similar manner. In Figure 1b, we also provide the agreement between all annotators. Upon completion of the annotation process, videos are assigned labels based on the majority agreement among participants.



(a) (b)

Figure 1: *Co-occurrence matrix of annotators' labels per class (a), confusion matrix showing agreement among annotators (b).*

3. Approach

After sourcing and preparing DEFCOMM-DB (cf. Section 2), we conduct a preprocessing step for both the audio and text modalities (cf. Section 3.1). Subsequently, we extract deep representations and expert-designed features from each modality (cf. Section 3.2). Using the obtained representations, we train machine learning models and fuse the best-performing models to check for their complementarity (cf. Section 4).

3.1. Preprocessing

For audio pre-processing, we first conduct speaker diarisation using the open-source *pyannote.audio* toolkit [26]. Afterwards, we normalise all audio tracks to ensure consistent loudness and convert them to 16 kHz and mono/16 bit. As for text, the provided subtitles come without any punctuation. Hence, we apply punctuation restoration [27] to segment each video's subtitles into sentences. In case there are several subjects in a video, we only keep the audio track and transcription of the person that is

communicating defensively in the respective situation such that every data point corresponds to exactly one person.

3.2. Feature Extraction

We utilise a set of open-source deep learning methods for representation learning and feature extraction from the audio (cf. Sections 3.2.1 to 3.2.4) and text (cf. Sections 3.2.5 and 3.2.6) modalities of the recordings in DEFCOMM-DB.

3.2.1. wav2vec2

Wav2Vec 2.0 [28] is a Transformer model pretrained for automatic speech recognition which has successfully been applied to speech emotion recognition [29]. We utilise a *base* version with 12 layers and about 95M parameters, trained on 960 hours of data from the Librispeech [30] dataset[2]. We extract 768-dimensional embeddings from the audio files by passing them to the pretrained model and averaging the resulting representations of its final layer.

3.2.2. auDeep

AUDEEP [31, 32] is employed for unsupervised representation learning from audio data. We extract Mel-spectrograms (128 Mels and Hanning window of width 80 ms with 50 % overlap) from raw waveforms in the dataset, followed by clipping power levels below four given thresholds $\{-30, -45, -60, -75\}$ dB to eliminate background noise. Next, distinct Recurrent Neural Network (RNN) autoencoders (2 hidden Gated Recurrent Unit (GRU) layers each with 256 units, unidirectional encoder, bidirectional decoder) are trained for 64 epochs with a batch size of 16, a learning rate of 0.001 and a keep probability of 80 % on each of these sets of spectrograms, and the learnt representations of each spectrogram are extracted as feature vectors of size 1024. Finally, these feature vectors are concatenated to form the final feature vector of size 4096.

3.2.3. DeepSpectrum

Additionally, DEEPSPECTRUM is applied to obtain deep audio representations utilising pre-trained Convolutional Neural Networks (CNNs) [33]. First, audio signals are transformed into Mel-spectrogram plots (*viridis* colour map, 64 Mels and Hanning window of width 32 ms with 50 % overlap). Using a sliding window of 1 s and overlap of 500 ms, the generated spectrograms are then forwarded through a DENSENET121 architecture (weights pre-trained on ImageNet), and the activations of the penultimate fully connected layer of the network are extracted, resulting in a 1024 dimensional feature set.

3.2.4. eGeMAPS

The expert-designed extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set [34] comprises 88 speech-related paralinguistic features and has been successfully applied in speech and affective computing tasks before [35, 36]. We extract the eGeMAPS features for every 2 s audio segment via OPENSMILE [37] using the standard configuration with a frame size of 1 s and a step size of 500 ms.

3.2.5. ELECTRA

We select a pretrained ELECTRA model [38], which has been used successfully in sentiment analysis [38] and emotion recognition [39]. Specifically, we use the *base* version with 12 layers[3] to generate 768-dimensional sentence embeddings by extracting the final layer's representation of the special `CLS` token.

Table 2: *Label distribution of each class in the speaker-independent training, development, and test partitions.*

Partition	Attack	Flight	Greater	Smaller	Total	# Clips
Train	44.1 %	20.7 %	17.0 %	18.1 %	72.0 %	188
Devel	37.8 %	27.0 %	18.9 %	16.2 %	14.2 %	37
Test	41.7 %	22.2 %	16.7 %	19.4 %	13.8 %	36

3.2.6. Semantic Similarity

To obtain more interpretable textual features, we opt for 2 reference sentences for each of the 4 classes: *"I am attacking you"* and *"You are not as great as you think"* for the class *Attack*, *"I am ending this conversation"* and *"I can not compete with you"* for the class *Flight*, *"I am superior to you"* and *"I am as great as you, if not greater"* for the class *Greater*, and *"I surrender to you"* and *"Please do not hurt me"* for the class *Smaller*. These sentences were inspired by Birkenbihl's definition of defensive communication [21]. We then utilise 4 pretrained sentence embedding models (*all-mpnet-base-v2, all-distilroberta-v1, all-MiniLM-L12-v2, MiniLM-L6-H384-uncased*) from the SBERT framework [40] to compute cosine distances between the representations of all sentences in a video and these 8 reference sentences. Each video is represented by a 32-dimensional feature vector computed through a weighted sum of sentence cosine distances, with weights based on sentence token count.

4. Experimental Settings

Partitioning: To reduce model bias toward specific speakers, we implement a speaker-independent partitioning strategy and divide the data into three training, development, and test sets. We further ensure that each set contains an approximately equal number of data points for each class (cf. Table 2).

Model training: We use Support Vector Machines (SVMs) with UAR as the evaluation metric. During the training, we optimise each SVM model on the development data split using the grid search hyperparameter optimisation and test the best-performing models on the unseen test partition. Specifically, we search for the optimal value of SVM's cost parameter (C) in the range of between 10^{-7} and 10^{1} on a logarithmic scale. Additionally, we test linear, Radial Basis Function (RBF), Sigmoid, and polynomial kernels. For the polynomial kernel, we further search for the optimal degree within the range of 2, 3, 4, and 5. Regarding the features extracted per segment, a majority voting over all segment-level predictions is conducted to obtain the class prediction for a video. Since the dataset has a slight imbalance in class distribution, we further apply the Synthetic Minority Over-sampling Technique (SMOTE) [41] with its default parameters (*random_state=seed_of_the_current_experiment, k_neighbors=5*) to the extracted features to address this issue. Every experiment is repeated with five fixed seeds.

Model fusion: Subsequently, we apply a simple late fusion approach to the trained models, where each model's class prediction is weighted according to the model's UAR on the development set. We experiment with the late fusion of both the audio-based models only and all models.

[2] https://huggingface.co/facebook/wav2vec2-base-960h

[3] https://huggingface.co/google/electra-base-discriminator

Table 3: *Overall and class-wise performance of the SVM models trained on clip-level features. Every experiment is repeated with five fixed seeds. The mean results over five seeds are provided. The standard deviation is given in parentheses. The best overall and class-wise performance on the test set for each modality is boldfaced. The best scores on the test set across all modalities (incl. late fusion) are marked with light grey shading. The chance-level is* 25.0 % *Unweighted Average Recall (UAR).*

	Overall [% UAR ↑]		**Attack [% Recall ↑]**		**Flight [% Recall ↑]**		**Greater [% Recall ↑]**		**Smaller [% Recall ↑]**	
Method	dev	test	dev	test	dev	test	dev	test	dev	test
Audio Modality										
w2v-base-960h	51.1 (± 0.0)	43.7 (± 0.0)	57.1 (± 0.0)	46.7 (± 0.0)	40.0 (± 0.0)	37.5 (± 0.0)	57.1 (± 0.0)	33.3 (± 0.0)	50.0 (± 0.0)	**57.1 (± 0.0)**
AUDEEP (fuse)	44.7 (± 2.9)	43.6 (± 4.3)	65.7 (± 5.4)	**73.3 (± 9.4)**	66.0 (± 4.9)	40.0 (± 9.4)	17.1 (± 16.7)	26.7 (± 13.3)	30.0 (± 12.5)	34.3 (± 21.4)
DS (DENSENET121)	49.8 (± 0.7)	**49.4 (± 1.5)**	77.1 (± 2.9)	66.7 (± 0.0)	60.0 (± 0.0)	**40.0 (± 5.0)**	28.6 (± 0.0)	36.7 (± 6.7)	33.3 (± 0.0)	54.3 (± 5.7)
eGeMAPS	55.0 (± 0.0)	47.6 (± 0.0)	57.1 (± 0.0)	60.0 (± 0.0)	70.0 (± 0.0)	37.5 (± 0.0)	42.9 (± 0.0)	**50.0 (± 0.0)**	50.0 (± 0.0)	42.9 (± 0.0)
Text Modality										
ELECTRA	54.6 (± 1.2)	43.2 (± 6.5)	72.9 (± 11.4)	**44.0 (± 6.8)**	56.0 (± 4.9)	35.0 (± 9.4)	42.9 (± 9.0)	56.7 (± 8.2)	46.7 (± 6.7)	37.1 (± 19.4)
Semantic Similarity	65.5 (± 0.7)	**52.2 (± 2.0)**	52.2 (± 3.5)	36.0 (± 6.8)	54.0 (± 4.9)	**50.0 (± 7.9)**	88.6 (± 5.7)	**80.0 (± 6.7)**	66.7 (± 0.0)	**42.9 (± 0.0)**
Model Fusion										
All audio models	58.9 (± 1.9)	52.0 (± 4.0)	72.9 (± 2.9)	72.0 (± 5.0)	76.0 (± 4.9)	45.0 (± 6.1)	40.0 (± 5.7)	36.7 (± 6.7)	46.7 (± 6.7)	54.3 (± 5.7)
All audio & text models	60.8 (± 3.0)	46.5 (± 3.3)	84.3 (± 7.0)	**73.3 (± 6.0)**	68.0 (± 4.0)	25.0 (± 7.9)	34.3 (± 7.0)	33.3 (± 18.3)	56.7 (± 8.2)	54.3 (± 5.7)

5. Results

Table 3 shows the classification results achieved by all evaluated models and their intra- and cross-modality late fusions. We report the means and standard deviations of the per-class recalls and the overall UAR computed over the five seeds for each experimental configuration. The overall best result can be found with a text-based model trained on semantic similarity features, achieving 52.2 % UAR for classifying the four types of defensive communication on the test set. In comparison, the best audio model utilising DEEPSPECTRUM features performs slightly worse at 49.4 % UAR.

However, the superiority of the text modality over the audio modality cannot be observed globally. For one, ELECTRA features only perform on par with the worst audio-based model at 43.2 % UAR. Furthermore, per-class recalls reveal that the semantic similarity approach owes its performance advantage to substantially higher detection rates of the *Flight* and *Greater* classes compared to all other models. The strategy of making oneself greater is detected with 80.0 % whereas both the ELECTRA (text) and eGeMAPs (audio) model fall behind at recalls of 56.7 % and 50.0 %, respectively. On the other hand, an active and confrontational strategy (*Attack*) is consistently detected better via the audio modality, with a top recall of 73.3 %, compared to 36.0 % achieved by the semantic similarity approach. From an affective computing point of view, this strategy will very likely coincide with higher degrees of arousal which can be detected from acoustic parameters such as increased loudness and pitch.

Finally, fusing predictions of all audio models increases the performance of this modality to be on par with the semantic similarity model at 52.0 %. However, adding the predictions obtained from the text-based models to the late fusion has a negative impact, dropping the overall UAR to 46.5 %. Nevertheless, based on our observations on per-class performance differences between audio and text-based models, the complementarity of the two modalities should be investigated further, e. g., by utilising different and more sophisticated fusion and multi-modal learning strategies.

6. Limitations and Discussion

Limitations of the current study can be found in the quantity and quality of the collected dataset. For one, certain types of defensive communication were more prevalent than others, resulting in a class imbalance. Furthermore, high-quality subtitles were not consistently available – a data-related issue which could be addressed by manual transcriptions. Lastly, some videos contained more than one defensive reaction at once, leading to difficulties in the labelling process. In the future, we plan to improve upon this issue by applying a multi-label approach and further creating a robust set of rules that can help to better distinguish the different classes.

7. Conclusions and Future Work

In the presented study, we investigated the feasibility of using machine learning to automatically detect four forms of defensive communication based on a popular model proposed by Birkenbihl [21]. For this purpose, we collected and annotated a database of 261 'real-world' video samples containing defensive communication from YouTube. We then evaluated an assortment of audio and text-based classification systems, utilising current state-of-the-art feature representations. Results showed that the four considered types of defensive communication could be automatically classified with a UAR of up to 52.2 %. Moreover, we observed that the four communication types differ in how distinctly they manifest across modalities.

In addition to addressing the limitations outlined in Section 6, future work should go into a deeper analysis of the underlying psychological, social and environmental factors such as emotional state, character traits, intentions, and relations between conversation partners that contribute to defensive communication. Apart from that, other classifications of defensive communication, such as the widely supported framework proposed by Gibb [5], should be adopted and compared.

Finally, we aim to make our algorithm suitable for deployment to embedded devices, utilising different techniques for model compression and hardware acceleration [42].

8. References

[1] S. Amiriparian, J. Han, M. Schmitt, A. Baird, A. Mallol-Ragolta, M. Milling, M. Gerczuk, and B. Schuller, "Synchronization in

interpersonal speech," *Frontiers in Robotics and AI*, vol. 6, p. 116, 2019.

[2] R. D. Jenifer and G. Raman, "Cross-cultural communication barriers in the workplace," *International Journal of Management*, vol. 6, no. 1, pp. 348–351, 2015.

[3] S. S. King, "The relationship between stress and communication in the organizational context," *Communication Studies*, vol. 37, no. 1, pp. 27–35, 1986.

[4] J. W. Pfeiffer, "Conditions that hinder effective communication," *The Pfeiffer Library*, vol. 6, no. 2, pp. 400–456, 1998.

[5] J. R. Gibb, "Defensive communication," *Journal of communication*, vol. 11, no. 3, pp. 141–148, 1961.

[6] P. Cramer, *Protecting the self: Defense mechanisms in action*. Guilford Press, 2006.

[7] G. H. Stamp, A. L. Vangelisti, and J. A. Daly, "The creation of defensiveness in social interaction," *Communication Quarterly*, vol. 40, no. 2, pp. 177–190, 1992.

[8] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychological inquiry*, vol. 1, no. 4, pp. 285–293, 1990.

[9] J. M. Gottman, "A theory of marital dissolution and stability." *Journal of family psychology*, vol. 7, no. 1, p. 57, 1993.

[10] J. A. Becker, B. Ellevold, and G. H. Stamp, "The creation of defensiveness in social interaction ii: A model of defensive communication among romantic couples," *Communication Monographs*, vol. 75, no. 1, pp. 86–110, 2008.

[11] J. A. Becker, J. R. Halbesleben, and H. Dan O'Hair, "Defensive communication and burnout in the workplace: The mediating role of leader–member exchange," *Communication Research Reports*, vol. 22, no. 2, pp. 143–150, 2005.

[12] J. C. Perry and F. F. Lanni, "Observer-rated measures of defense mechanisms," *Journal of personality*, vol. 66, no. 6, pp. 993–1024, 1998.

[13] P. Lerner and H. Lerner, "Rorschach assessment of primitive defenses in borderline personality structure," *Borderline phenomena and the Rorschach test*, pp. 257–274, 1980.

[14] S. H. Cooper, J. C. Perry, and D. Arnow, "An empirical approach to the study of defense mechanisms: I. reliability and preliminary validity of the rorschach defense scales," *Journal of Personality Assessment*, vol. 52, no. 2, pp. 187–203, 1988.

[15] G. E. Vaillant, "Ego mechanisms of defense and personality psychopathology." *Journal of abnormal psychology*, vol. 103, no. 1, p. 44, 1994.

[16] J. C. Perry and M. Henry, "Studying defense mechanisms in psychotherapy using the defense mechanism rating scales," *Defense mechanisms: Theoretical, research and clinical perspectives*, vol. 136, pp. 165–186, 2004.

[17] M. Di Giuseppe and J. C. Perry, "The hierarchy of defense mechanisms: assessing defensive functioning with the defense mechanisms rating scales q-sort," *Frontiers in Psychology*, vol. 12, 2021.

[18] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. Interspeech*, 2012, pp. 254–257.

[19] L. Batrinca, B. Lepri, N. Mana, and F. Pianesi, "Multimodal recognition of personality traits in human-computer collaborative tasks," in *Proc. ICMI*, 2012, pp. 39–46.

[20] M. Askari, S. B. M. Noah, S. A. B. Hassan, and M. B. Baba, "Comparison the effects of communication and conflict resolution skills training on marital satisfaction," *International Journal of Psychological Studies*, vol. 4, no. 1, p. 182, 2012.

[21] V. F. Birkenbihl, *Kommunikationstraining: zwischenmenschliche Beziehungen erfolgreich gestalten*. mvg Verlag, 2013.

[22] J. Parker and E. Coiera, "Improving clinical communication: a view from psychology," *Journal of the American Medical Informatics Association*, vol. 7, no. 5, pp. 453–461, 2000.

[23] M. L. Friedlander and G. S. Schwartz, "Toward a theory of strategic self-presentation in counseling and psychotherapy." *Journal of Counseling Psychology*, vol. 32, no. 4, p. 483, 1985.

[24] J. M. Civikly, R. Wayne Pace, and R. M. Krause, "Interviewer and client behaviors in supportive and defensive interviews," *Annals of the International Communication Association*, vol. 1, no. 1, pp. 347–361, 1977.

[25] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[26] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, Brno, Czech Republic, August 2021.

[27] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme, "Fullstop: Multilingual deep models for punctuation prediction," in *Proc. SwissText*. Winterthur, Switzerland: CEUR, June 2021.

[28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[29] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[31] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio," in *Proc. of DCASE 2017*. Munich, Germany: IEEE, November 2017, pp. 17–21.

[32] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks," *JMLR*, vol. 18, pp. 1–5, April 2018.

[33] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. Interspeech*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 3512–3516.

[34] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "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, 2015.

[35] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?" in *Proc. Interspeech*, 2018, pp. 147–151.

[36] S. Amiriparian, "Deep representation learning techniques for audio signal processing," Ph.D. dissertation, Technische Universität München, 2019.

[37] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*. Firenze, Italy: ACM, 2010, pp. 1459–1462.

[38] K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: pre-training text encoders as discriminators rather than generators," in *Proc. ICLR*, Addis Ababa, Ethiopia, 2020.

[39] L. Christ, S. Amiriparian, M. Milling, I. Aslan, and B. W. Schuller, "Automatic emotion modelling in written stories," *arXiv preprint arXiv:2212.11382*, 2022.

[40] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP*. Hong Kong, China: ACL, Nov. 2019, pp. 3982–3992.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[42] S. Amiriparian, T. Hübner, V. Karas, M. Gerczuk, S. Ottl, and B. W. Schuller, "Deepspectrumlite: A power-efficient transfer learning framework for embedded speech and audio processing from decentralized data," *Frontiers in AI*, vol. 5, 2022.