# Computer Audition for Emotional Wellbeing

Inaugural-Dissertation
zur Erlangung des Doktorgrades (Dr. rer. nat.)
an der
Fakultät für Angewandte Informatik
der Universität Augsburg

vorgelegt von

**Alice Baird**

2021

| | |
|---|---|
| **Erstgutachter:** | Prof. Dr. Björn Schuller |
| **Zweitgutachter:** | Prof. Dr. Elisabeth André |
| **Drittgutachter:** | PD Dr.- Ing. habil. Ronald Böck |
| **Tag der mündlichen Prüfung:** | 20.10.2022 |

# Acknowledgements

Throughout my doctoral research, I have been fortunate to work with many talented researchers from various scientific disciplines. As such, there is a long list of individuals that I am grateful to for their positive impact on this thesis. Firstly, I would like to thank my supervisor, Dr. Björn Schuller, for his support and guidance. The opportunities I have had throughout my time in Germany, under his supervision, have been inspiring and allowed me to develop a research practice that I am excited to continue to grow. I would also like to thank my co-authors and collaborators with whom I have had many knowledge exchanges, discussing strategies that have greatly benefited this research. I am particularly grateful to my colleague Lukas Stappen for his collaboration over the years and for taking the time to comment on this thesis. Further thanks also go to Dr. Shahin Amiriparian and Manuel Milling, who have both been great collaborators and offered meaningful feedback on this work. Furthermore, a great deal of gratitude goes to Dr. Anton Batliner, Dr. Emilia Parada-Cabalerio, Dr. Eduardo Coutinho, Dr. Nicholas Cummins, Dr. Julia Hirshberg, Dr. Stina Jørgensen, Dr. Gil Keren, Dr. Eva-Maria Meßner, Dr. Panagiotis Tzirakis, Dr. Kun Qian, Lukas Christ, Alexander Gebhard, Shuo Liu, Adria Mallol-Ragolta, Silvan Mertes, Zhao Ren, Georgios Rizos, Maximilian Schmitt, Meishu Song, Andreas Triantafyllopoulos, Thomas Wiest, and Sandra Zänkert.

I have been fortunate to have received funding to complete my doctoral research from numerous sources. These include the EU Horizon 2020 DE-ENIGMA project, the Centre of Digitisation Bavaria, and the Reinhart Konsolleck DFG AUDION0MOUS project.

Lastly, I would like to highlight the personal support I have received from my loving family and friends throughout the development of this research. I am fully aware that I would not have reached this point without them and am enormously grateful. Thank you.

**Alice Baird**

October 2021

# Abstract

This thesis is focused on the application of computer audition (i. e., machine listening) methodologies for monitoring states of emotional wellbeing. Computer audition is a growing field and has been successfully applied to an array of use cases in recent years. There are several advantages to audio-based computational analysis; for example, audio can be recorded non-invasively, stored economically, and can capture rich information on happenings in a given environment, e. g., human behaviour. With this in mind, maintaining emotional wellbeing is a challenge for humans and emotion-altering conditions, including stress and anxiety, have become increasingly common in recent years. Such conditions manifest in the body, inherently changing how we express ourselves. Research shows these alterations are perceivable within vocalisation, suggesting that speech-based audio monitoring may be valuable for developing artificially intelligent systems that target improved wellbeing. Furthermore, computer audition applies machine learning and other computational techniques to audio understanding, and so by combining computer audition with applications in the domain of computational paralinguistics and emotional wellbeing, this research concerns the broader field of empathy for Artificial Intelligence (AI). To this end, speech-based audio modelling that incorporates and understands paralinguistic wellbeing-related states may be a vital cornerstone for improving the degree of empathy that an artificial intelligence has.

To summarise, this thesis investigates the extent to which speech-based computer audition methodologies can be utilised to understand human emotional wellbeing. A fundamental background on the fields in question as they pertain to emotional wellbeing is first presented, followed by an outline of the applied audio-based methodologies. Next, detail is provided for several machine learning experiments focused on emotional wellbeing applications, including analysis and recognition of under-researched phenomena in speech, e. g., anxiety, and markers of stress. Core contributions from this thesis include the collection of several related datasets, hybrid fusion strategies for an emotional gold standard, novel machine learning strategies for data interpretation, and an in-depth acoustic-based computational evaluation of several human states. All of these contributions focus on ascertaining the advantage of audio in the context of modelling emotional wellbeing. Given the sensitive nature of human wellbeing, the ethical implications involved with developing and applying such systems are discussed throughout.

# Contents

# List of Publications

During the development of this doctoral research, the author has (co-)authored several peer-reviewed journal articles (18) and conference proceedings (34), some of which will be detailed within this thesis. The following is an exhaustive list of these publications given in reverse chronological order.


[1] A. **Baird**, S. Mertes, M. Milling, L. Stappen, T. Wiest, E. André, and B. W. Schuller, "A prototypical network approach for evaluating generated emotional speech," in *Proceedings of INTERSPEECH, 22nd Annual Conference of the International Speech Communication Association*, (Brno, Czechia), pp. 3161–3165, ISCA, 2021.

[2] A. **Baird**, L. Stappen, L. Christ, L. Schumann, E. Messner, and B. W. Schuller, "A physiologically-adapted gold standard for arousal during stress," in *Proceedings of MuSe'21, 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Chengdu, China), ACM, 2021.

[3] A. **Baird**, S. Amiriparian, M. Milling, and B. W. Schuller, "Emotion recognition in public speaking scenarios utilising an LSTM-RNN approach with attention," in *Proceedings of SLT 2021, IEEE Spoken Language Technology Workshop*, pp. 397–402, IEEE, 2021.

[4] A. **Baird**, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, B. M. Kudielka, S. Sturmbauer, N. Rohleder, E.-M. Messner, and B. Schuller, "Evaluating speech-based recognition of emotional & physiological markers of stress," *Frontiers in Computer Science, Human-Media Interaction*, vol. to appear, 2021.

[5] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. **Baird**, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. M. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates," in *Proceedings of INTERSPEECH, 22nd Annual Conference of the International Speech Communication Association*, (Brno, Czechia), pp. 431–435, ISCA, 2021.

[6] L. Stappen, A. **Baird**, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements," *IEEE Transactions on Affective Computing*, no. 1, pp. 1–16, 2021.

[7] H. Coppock, A. Gaskell, P. Tzirakis, A. **Baird**, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: A pilot study," *BMJ Innovations*, vol. 7, no. 2, pp. 356–362, 2021.

[8] L. Stappen, A. **Baird**, L. Christ, L. Schumann, B. Sertolli, E. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proceedings of MuSe'21, 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Chengdu, China), ACM, 2021.

[9] L. Stappen, L. Schumann, B. Sertolli, A. **Baird**, B. Weigel, E. Cambria, and B. W. Schuller, "MuSe-Toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox," in *Proceedings of MuSe'21, 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Chengdu, China), ACM, 2021.

[10] B. Schuller, A. **Baird**, A. Gebhard, S. Amiriparian, G. Keren, M. Schmitt, and N. Cummins, "New avenues in audio intelligence: Towards holistic real-life audio understanding," *Trends in Hearing*, vol. to appear, 2021.

[11] L. Stappen, A. **Baird**, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intelligent Systems*, vol. 36, no. 2, pp. 88–98, 2021.

[12] A. **Baird**, N. Cummins, S. Schnieder, and B. W. Schuller, "An evaluation of the effect of anxiety on speech – computational prediction of anxiety from sustained vowels," in *Proceedings of INTERSPEECH, 21st Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 4951–4955, ISCA, 2020.

[13] A. **Baird** and B. Schuller, "Considerations for a more ethical approach to data in AI: On data representation and infrastructure," *Frontiers in Big Data*, vol. 3, 2020. Art. no. 25.

[14] A. **Baird**, M. Song, and B. Schuller, "Interaction with the soundscape: Exploring emotional audio generation for improved individual wellbeing," in *Artificial Intelligence*

*in HCI* (H. Degen and L. Reinerman-Jones, eds.), (Cham, Switzerland), pp. 229–242, Springer, 2020.

[15] S. Amiriparian, M. Gerczuk, S. Ottl, L. Stappen, A. **Baird**, L. Koebe, and B. Schuller, "Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, 2020. Art. no. 19.

[16] A. D. MacIntyre, G. Rizos, A. Batliner, A. **Baird**, S. Amiriparian, A. Hamilton, and B. W. Schuller, "Deep attentive end-to-end continuous breath sensing from speech," in *Proceedings of INTERSPEECH, 21st Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 2082–2086, ISCA, 2020.

[17] S. Mertes, A. **Baird**, D. Schiller, B. W. Schuller, and E. André, "An evolutionary-based generative approach for audio data augmentation," in *Proceedings of MMSP 2020, IEEE 22nd International Workshop on Multimedia Signal Processing*, (Tampere, Finland), IEEE, 2020.

[18] E. Parada-Cabaleiro, A. Batliner, A. **Baird**, and B. Schuller, "The perception of emotional cues by children in artificial background noise," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 169–182, 2020.

[19] G. Rizos, A. **Baird**, M. Elliott, and B. Schuller, "StarGAN for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *Proceedings of ICASSP 2020, International Conference on Acoustics, Speech and Signal Processing*, (Barcelona, Spain), pp. 3502–3506, IEEE, 2020.

[20] B. W. 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," in *Proceedings of INTERSPEECH, 21st Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 2042–2046, ISCA, 2020.

[21] L. Stappen, A. **Baird**, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, "MuSe 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *Proceedings of MuSe'20, 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Seattle, WA, USA), pp. 35–44, ACM, 2020.

[22] A. **Baird**, S. Amiriparian, M. Berschneider, M. Schmitt, and B. Schuller, "Predicting biological signals from speech: Introducing a novel multimodal dataset and results," in *Proceedings of MMSP 2019, 21st International Workshop on Multimedia Signal Processing*, (Kuala Lumpur, Malaysia), IEEE, 2019.

[23] A. **Baird**, S. Amiriparian, N. Cummins, S. Sturmbauer, J. Janson, E.-M. Messner, H. Baumeister, N. Rohleder, and B. Schuller, "Using speech to predict sequentially measured cortisol levels during a Trier Social Stress Test," in *Proceedings of INTERSPEECH, 20th Annual Conference of the International Speech Communication Association*, (Graz, Austria), pp. 534–538, ISCA, 2019.

[24] A. **Baird**, S. Amiriparian, and B. Schuller, "Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation," in *Proceedings of MMSP 2019, 21st International Workshop on Multimedia Signal Processing*, (Kuala Lumpur, Malaysia), IEEE, 2019.

[25] A. **Baird**, E. Coutinho, J. Hirschberg, and B. W. Schuller, "Sincerity in acted speech: Presenting the sincere apology corpus and results," in *Proceedings of INTERSPEECH, 20th Annual Conference of the International Speech Communication Association*, (Graz, Austria), pp. 539–543, ISCA, 2019.

[26] S. Amiriparian, A. Awad, M. Gerczuk, L. Stappen, A. **Baird**, S. Ottl, and B. Schuller, "Audio-based recognition of bipolar disorder utilising capsule networks," in *Proceedings of IJCNN 2019, International Joint Conference on Neural Networks*, (Budapest, Hungary), IEEE, 2019. Art. no. 19242.

[27] S. Amiriparian, M. Gerczuk, E. Coutinho, A. **Baird**, S. Ottl, M. Milling, and B. Schuller, "Emotion and themes recognition in music utilising convolutional and recurrent neural networks," in *Proceedings of the MediaEval 2019 Multimedia Benchmark Workshop*, vol. 2670, (Sophia Antipolis, France), 2019.

[28] 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, 2019. Art. no. 116.

[29] F. Dong, K. Qian, Z. Ren, A. **Baird**, X. Li, Z. Dai, B. Dong, F. Metze, Y. Yamamoto, and B. W. Schuller, "Machine listening for heart status monitoring: Introducing and benchmarking HSS – the Heart Sounds Shenzhen Corpus," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2082–2092, 2019.

[30] A. Mallol-Ragolta, M. Schmitt, A. **Baird**, N. Cummins, and B. Schuller, "Performance analysis of unimodal and multimodal models in valence-based empathy recognition," in *Proceedings of FG 2019, 14th IEEE International Conference on Automatic Face & Gesture Recognition*, (Lille, France), IEEE, 2019.

[31] E. Marchi, B. Schuller, A. **Baird**, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, P. Robinson, I. Davies, T. Baltrušaitis, A. Adams, M. Mahmoud, O. Golan, S. Fridenson-Hayo, S. Tal, S. Newman, N. Meir-Goren, A. Camurri, S. Piana, S. Bölte, M. Sezgin, N. Alyuz, A. Rynkiewicz, and A. Baranger, "The ASC-Inclusion perceptual serious gaming platform for autistic children," *IEEE Transactions on Games*, vol. 11, no. 4, pp. 328–339, 2019.

[32] B. Schuller, S. Amiriparian, G. Keren, A. **Baird**, M. Schmitt, and N. Cummins, "The next generation of audio intelligence: A survey-based perspective on improving audio analysis," in *Proceedings of ISAAR 2019, 7th International Symposium on Auditory and Audiological Research*, vol. 7, (Nyborg, Denmark), pp. 101–112, 2019.

[33] M. Song, Z. Yang, A. **Baird**, E. Parada-Cabaleiro, Z. Zhang, Z. Zhao, and B. Schuller, "Audiovisual analysis for recognising frustration during game-play: Introducing the multimodal game frustration database," in *Proceedings of ACII 2019, 8th International Conference on Affective Computing and Intelligent Interaction*, (Cambridge, UK), pp. 517–523, IEEE, 2019.

[34] Z. Yang, K. Qian, Z. Ren, A. **Baird**, Z. Zhang, and B. Schuller, "Learning multi-resolution representations for acoustic scene classification via neural networks," in *Proceedings of 7th Conference on Sound and Music Technology*, (Hei Long Jiang, China), pp. 133–143, 2019.

[35] A. **Baird**, S. H. Jørgensen, E. Parada-Cabaleiro, N. Cummins, S. Hantke, and B. Schuller, "The perception of vocal traits in synthesized voices: Age, gender, and human likeness," *Journal of the Audio Engineering Society*, vol. 66, no. 4, pp. 277–285, 2018.

[36] A. **Baird**, E. Parada-Cabaleiro, C. Fraser, S. Hantke, and B. Schuller, "The perceived emotion of isolated synthetic audio: The EmoSynth dataset and results," in *Proceedings of AM '18, Audio Mostly 2018 on Sound in Immersion and Emotion*, (Wrexham, UK), ACM, 2018. Art. no. 7.

[37] A. **Baird**, E. Parada-Cabaleiro, S. Hantke, F. Burkhardt, N. Cummins, and B. Schuller, "The perception and analysis of the likeability and human likeness of synthesized speech,"

in *Proceedings of INTERSPEECH, 19th Annual Conference of the International Speech Communication Association*, (Hyderabad, India), pp. 2863–2867, ISCA, 2018.

[38] S. Amiriparian, A. **Baird**, S. Julka, A. Alcorn, S. Ottl, S. Petrović, E. Ainger, N. Cummins, and B. Schuller, "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in *Proceedings of INTERSPEECH, 19th Annual Conference of the International Speech Communication Association*, (Hyderabad, India), pp. 2334–2338, ISCA, 2018.

[39] B. Schuller, S. Steidl, P. Marschik, H. Baumeister, F. Dong, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, **A. Baird**, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & self-assessed affect, crying & heart beats," in *Proceedings of INTERSPEECH, 19th Annual Conference of the International Speech Communication Association*, (Hyderabad, India), pp. 122–126, ISCA, 2018.

[40] N. Cummins, A. **Baird**, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[41] K. Grabowski, A. Rynkiewicz, A. Lassalle, S. Baron-Cohen, B. Schuller, N. Cummins, A. **Baird**, J. Podgórska-Bednarz, A. Pieniążek, and I. Łucka, "Emotional expression in psychiatric conditions: New technology for clinicians," *Psychiatry and Clinical Neurosciences*, vol. 73, no. 2, pp. 50–62, 2018.

[42] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. **Baird**, and B. W. Schuller, "Categorical vs dimensional perception of Italian emotional speech," in *Proceedings of INTERSPEECH, 19th Annual Conference of the International Speech Communication Association*, (Hyderabad, India), pp. 3638–3642, ISCA, 2018.

[43] K. Qian, C. Janott, Z. Zhang, J. Deng, A. **Baird**, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Teaching machines on snoring: A benchmark on computer audition for snore sound excitation localisation," *Archives of Acoustics*, vol. 43, no. 3, pp. 465–475, 2018.

[44] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. **Baird**, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 3, pp. 662–669, 2018.

[45] A. **Baird**, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proceedings of INTERSPEECH, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), pp. 849–853, ISCA, 2017.

[46] A. **Baird**, S. H. Jørgensen, E. Parada-Cabaleiro, S. Hantke, N. Cummins, and B. Schuller, "Perception of paralinguistic traits in synthesized voices," in *Proceedings of AM '17, 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, (London, UK), ACM, 2017. Art. no. 17.

[47] 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 *Proceedings of INTERSPEECH, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), pp. 3512–3516, ISCA, 2017.

[48] E. Parada-Cabaleiro, A. Batliner, A. **Baird**, and B. Schuller, "The SEILS dataset: Symbolically encoded scores in modern-early notation for computational musicology," in *Proceedings of ISMIR 2017, International Society for Music Information Retrieval Conference*, (Suzhou, China), pp. 575–581, 2017.

[49] E. Parada-Cabaleiro, A. **Baird**, A. Batliner, N. Cummins, S. Hantke, and B. Schuller, "The perception of emotions in noisified nonsense speech," in *Proceedings of INTERSPEECH, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), pp. 3246–3250, ISCA, 2017.

[50] E. Parada-Cabaleiro, A. **Baird**, N. Cummins, and B. W. Schuller, "Stimulation of psychological listener experiences by semi-automatically composed electroacoustic environments," in *Proceedings of ICME 2017, IEEE International Conference on Multimedia and Expo*, (Hong Kong, China), pp. 1051–1056, IEEE, 2017.

[51] K. Qian, Z. Zhang, A. **Baird**, and B. Schuller, "Active learning for bird sounds classification," *Acta Acustica united with Acustica*, vol. 103, no. 3, pp. 361–364, 2017.

[52] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. **Baird**, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, sincerity & native language," in *Proceedings of INTERSPEECH, 17th Annual Conference of the International Speech Communication Association*, (San Francisco, CA, USA), pp. 2001–2005, ISCA, 2016.

# Introduction

## 1.1 Motivation

Sound is an immersive, continuous, complex array of pressure waves. From speech alone, a deep understanding of a human's current state of being can be understood [1]. The field of acoustics is a branch of science that is concerned with the production, transmission, and effects of sound. Consistent with *Lindsay's Wheel of Acoustics*[1], this thesis focuses on the acoustic targets related to human-audible information. These are associated with the acoustic domains of the life sciences and the arts, i. e., speech, physiology, and psychology.

Computer (or machine) audition is a field of research focused on the understanding of audio by machines [2]. Research in computer audition applies methods from areas including signal processing, machine learning, and psychology. More specific technical topics of interest within computer audition include automatic speech recognition [3], speech enhancement [4], and acoustic event detection [5]. The field has continued to grow over many years and has been advanced by the rise in deep learning [6]. Furthermore, there are now several well-established methods and openly available toolkits which allow for the analysis and recognition of a variety of human states from representations of speech [7].

At a subconscious level, humans have a remarkable ability to interpret and process fine-grained and subtle changes in information transmitted via sound, including mood, intention, and type of activity [8]. This is a phenomenon that computer audition researchers have been attempting to replicate computationally with machine learning and signal processing techniques for many years [1, 8]. Within the machine learning community, various modalities can be modelled, including video, text, and biological signals. In comparison to these, audio has shown to be particularly dominant in regards to human behaviour modelling, e. g., concerning speech emotion recognition, where emotional activation is far better modelled with audio [9]. As such, audio is particularly suited to in-the-wild and non-invasive monitoring use cases, as occlusions that would be typical from the video modality are avoidable [10]. Specifically, throughout the last decade, the field of computational paralinguistics has shown great promise for robustly recognising several human states and focusing on how an individual speaks

[1] R. Bruce Lindsey, "Lindsay's Wheel of Acoustics", Journal of the Acoustical Society of America (1964).

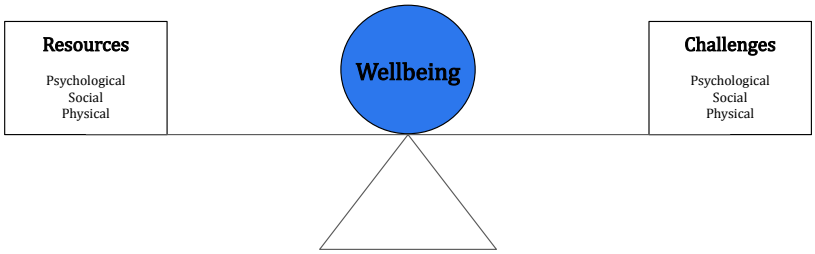



Figure 1.1: An adaptation of the "new definition for wellbeing" proposed by [12]. The figure illustrates how wellbeing is managed via the balancing act between the psychological, social or physical resources available and the challenges faced.

rather than explicitly what the individual has said. With this in mind, managing states of human wellbeing has vast benefits [11] for which speech-based approaches may be suited.

Wellbeing is a universal term that is complex to define but essentially refers to life satisfaction. A positive state of wellbeing can be said to be an equilibrium between current psychological, social and physical challenges and available resources [12] (see Figure 1.1 for a depiction). Managing the psychological aspect of wellbeing, described as emotional wellbeing, can be difficult for many individuals in modern society [13] due to unexpected challenges which exceed the limited resources available. This can result in negative emotional responses caused by conditions such as stress and anxiety. It is well established in research that the speech signal is affected directly by several physical and environmental changes [7], suggesting again that computational speech analysis frameworks can aid in the detection of key markers of emotional wellbeing, specifically as it pertains to stress and anxiety. Such conditions are currently under researched from a computational analysis perspective.

As well as audio, there have been many recent advancements for the computational analysis and representation of human wellbeing by machines coming from the computer vision domain [14]. Such progress is promising as combining modalities has shown to have great benefits [15]. As well as improving understanding of emotional wellbeing, the development of more empathic AI in general may be advanced. Typically within media, empathy in AI pertains to improving the degree of empathy which is present in human-centred AI interactions, e. g., by companion robotics or assistants. However, AI empathy is not limited to robotics and there is much discussion relating to higher-level design choices and lower-level machine reasoning being made with more consideration towards a user's emotional state and current needs, e. g., the time for which a notification is pushed on a smart device. As AI is becoming an integrated part of human life, reasoning with empathy along with consideration to the ethical concerns of this, should all be incorporated into a more positive AI ecosystem. In this regard, if computational audio approaches can be applied to

understand emotional wellbeing, this can also assist machine self-awareness and, in turn, the ability for an AI to respond with empathy.

With these aspects in mind, the core aim of this thesis is to facilitate research in computer audition on aspects of emotional wellbeing, particularly concerning paralinguistic speech. To summarise the key points from the above passage, the motivation for this thesis is four-fold:

1. **The ubiquitous and non-invasive nature of audio**. Audio is an immersive signal which allows for holistic recognition and analysis of several human states [8]. The advantages from a technical perspective include i) continuous real-time monitoring ii) pseudo-anonymous storage iii) low-resource computational expense .

2. **The advancements in machine learning for which audio can be applied**. Advancements in machine learning research mean that several open-source and highly relevant resources are available to build upon for computer audition [16] from neighbouring fields, e. g., natural language processing [17] and computer vision [18].

3. **The limited prior research and literature**. When it comes to the application of state-of-the-art computer audition-based machine learning methods for states of emotional wellbeing, there is limited previous research to build upon, particularly for mental health conditions such as anxiety [19]. This is particularly motivating given the drive for more empathic AI. A meaningful aspect of improving the ability for AI to develop empathy is understanding the current mood of the person interacting with the AI.

4. **The continued decrease in general wellbeing in modern societies**. Several aspects of daily life contribute to lower wellbeing, e. g., loneliness [20] or connection to nature [21], and it is estimated that one in five individuals may have a mental disorder. English-speaking regions have the highest lifetime prevalence [22]. Such a decline suggests that more research into assistive technologies is needed to aid individuals in retaining the balance between resources and challenges.

## 1.2 Research Questions

As described, the core focus of this thesis is to apply computational audio-based methodologies to a variety of novel applications targeted at monitoring emotional wellbeing. With this in mind, some formulated research questions (RQ) which this thesis explores include:

**RQ-1: To what extent can computer audition methodologies be harnessed to monitor states of emotional wellbeing? Furthermore, which aspects of emotional**

**wellbeing can not be suitably modelled by audio?** To explore this, a number of audio-based experiments are outlined which apply machine learning methodologies and acoustic analysis to several datasets including those which are based on the well-established Trier Social Stress Test (TSST).

**RQ-2: How does audio perform as a uni-modal signal compared with other modalities and multimodal fusion? Additionally, which feature representations of audio are useful for monitoring emotional wellbeing?** To evaluate this, a multimodal dataset that includes the modalities of audio, video, text, and physiological signals is utilised, and a series of experiments on individual and combined modalities will focus on the recognition of markers of stress. Within this context, a number of hand-crafted vs image-based audio features, which are known to be strong in the area of computational paralinguistics, are applied throughout all experiments.

**RQ-3: What data is available for academic research into the computational analysis of states of emotional wellbeing from audio-based signals, and how can concerns relating to data scarcity be tackled?** To gain an understanding of this, an overview of related datasets will be given in Chapter 2, as well as details for a number of datasets that have been collected in the context of this thesis Chapter 3. Furthermore, generative approaches and other conventional data augmentation is explored in the context of emotional speech.

**RQ-4: How can computational approaches, when applied to emotional wellbeing, benefit from interdisciplinary collaboration? Furthermore, how can outcomes be better understood (interpretable AI) and utilised by those working in related fields of research?** Given the sensitive nature of states of emotional wellbeing, the majority of conducted experiments were made in collaboration with the data owner, or with guidance from researchers with expert knowledge of the domains. Accordingly, communication within a number of experiments is made in an open manner with interpretations of specific results provided based on the literature from adjacent fields.

## 1.3 Contributions

This thesis provides many contributions and findings to the community. At a high-level, the main contributions include:

- A rigorous computational evaluation of under-researched areas of speech-based states of individual emotional wellbeing, e. g., anxiety [19], stress [23], and dimensions of

emotion in stress-induced scenarios [24, 25] – providing several novel findings, which are applicable to researchers in computational paralinguistics, affective computing, machine learning as well as the broader research community.

- The development of novel strategies for methodologies within computer audition applicable to the machine learning domain, including data interpretation [26] and fusion and adaptation of subjective signals [24, 27] - with frameworks made open-source to the community where appropriate.
- A series of interdisciplinary collaborations have been conducted, which support accessibility and the application of computer science methods in the area of human emotional wellbeing – validating a number of machine learning approaches with strong psychological backing and motivation on the experimental design, for example, in [19, 23, 24].
- Several (more than five) uni- and multi-modal datasets in the domain of individual emotional wellbeing have been acquired, processed, and evaluated during the period of this doctoral research, with many available in the public research domain [28, 25, 29–32]. An overview of these, as they pertain to the thesis, is given in Chapter 3, Table 3.1.

## 1.4 Thesis Structure

This thesis is structured as follows:

- **Chapter 2 (Background):** This chapter will offer a general introduction and overview for several fields of research that relate to the topics of this thesis, namely, computer audition, computational paralinguistics, and an overview of modelling emotion from speech as it pertains to both empathy in artificial intelligence and the field of speech emotion recognition in general.
- **Chapter 3 (Methodology):** In this chapter, more specific detail is given for the methodologies which are applied in Chapter 4 for the understanding of emotional wellbeing. This chapter gives detail on the data acquisition process for the datasets utilised, the theoretical detail for audio representations, and the applied machine learning architectures for both recognition and generation of audio.
- **Chapter 4 (Experiments):** This chapter will offer a series of experiments that evaluate computer audition's efficacy in the context of emotional wellbeing. The experiments within this chapter address each of the above-mentioned research questions in various

ways, and the methodologies described in Chapter 3 will be proposed for application in several emotional wellbeing-related targets. In general, these experiments should be considered an anchor for any conclusions made from this thesis.

- **Chapter 5 (Concluding remarks):** This chapter will summarise the findings on the main topics explored within this thesis, relating particular findings to each of the core research questions. Furthermore, a discussion is provided for the ethical aspects that are found to be relevant and essential for this area of research, with detail on the limitations of the current work and an outlook for this area of research in general.

# Background

This chapter introduces the higher-level background and essential concepts for the core topics of this thesis, namely, Computer Audition, Computational Paralinguistics and Empathy in Artificial Intelligence as it relates to Speech Emotion Recognition. The deeper theoretical background and more technical aspects will be introduced in the preceding section (see Chapter 3) as it pertains to the conducted experiments of Chapter 4. As the literature in these areas is extensive, in the following section – unless discussing fundamentals – the scope of all discussion is limited to works which relate only to emotional wellbeing adjacent applications. With this in mind, a brief definition of the topic will be given at the beginning of each section.

## 2.1 Computer Audition in General

Computer audition is a discipline of computer science and engineering that has been gaining in popularity throughout the last couple of decades [33, 34]. Where Audition refers to the Latin verb *audire*, meaning 'to hear', essentially, Computer Audition is the computational understanding of acoustic information by machines and can cover a wide range of computational tasks, to both analyse a human phenomenon in nature as well generate human or natural audio expression. As with many fields which are inherently interdisciplinary and less deeply established, computer audition has several alternative names, including Machine Audition [35], Intelligent Audio Analysis [36], and Audio Information Retrieval [37]. With the idea that computer audition can be concerned with any acoustic signal, as mentioned earlier based on the breakdown provided by Lindsey's Wheel of Acoustics, this thesis is concerned primarily with the acoustics derived from vocalisations, and in that regard crosses between the life sciences and the arts.

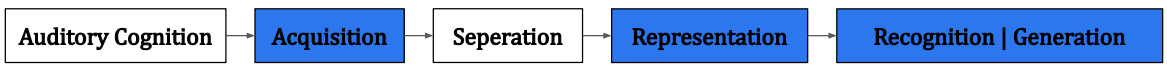



Figure 2.1: An overview of the topics in the field of computer audition, and how they are related to one another within a research pipeline. Where the topic is highlighted in blue, this represents computer audition problems discussed and tackled within this thesis.

There are several sub-problems that researchers are interested in for the field of computer audition, as shown in Figure 2.1, these include auditory cognition [38], audio separation and enhancement [39], audio representations [40], audio sequence modelling [41], and audio generation [42]. At the earliest point in a computer audition pipeline, a fundamental understanding of aspects of auditory cognition for a particular phenomenon is needed. Auditory cognition refers to the human brain's process of determining meaning and deciding upon action based on the current soundscape, i. e., how is a sound perceived and the mechanical process performed by the auditory cortex of the brain.

Auditory cognition is meaningful to computer audition as through understanding how sound is perceived. Engineers and computer scientists can – particularly now with machine learning – be more mindful and fine-tune specific parameters based on this. Current computer audition researchers continue to utilise the already extremely well-established research in this area, particularly as it pertains to the mechanics of the human ear [43], and perception, e. g., perception of loudness [44]. However, research for the more subjective sonic interactions, e. g., emotion in speech, and how the brain understands this, continues to be researched, e. g., how emotional state can effect loudness [45].

With this in mind, applying the cognitive sciences to specific phenomena in audio during computational analysis and generation may assist in ethical developmental choices. For example, in relation to bias, a higher-pitched voice is often perceived to have less authority and therefore lower-pitched more masculine voices are often more successfully able to convey a message [46]. Selection bias (i. e., unbalanced population demographics) then also becomes a risk in the context of machine learning as models may learn to bias underrepresented vocal characteristics in this same way, and researchers must be aware of such cognitive-derived imbalances [47]. Furthermore, cognition is extremely subject-dependent, and it is known that several environmental aspects e. g., degree of noise, will affect how an individual perceives the emotion speech [48].

### 2.1.1 Computer Audition and Machine Learning

In today's computer audition landscape, machine learning is a core resource, and is applied for the analysis and generation of various human phenomena in the experiments conducted within this thesis. With this in mind, it is essential to give a fundamental description of the typical audio-based machine learning pipeline. In Figure 2.2, an illustration of a feature-based machine learning pipeline is depicted (above) along with an end-to-end approach where feature extraction is omitted, and data processing is less extensive (below). It is important to note that although fruitful, the end-to-end approach can be a challenge when data is sparse [49]. For analysis of speech-driven computer audition aspects, feature-based machine

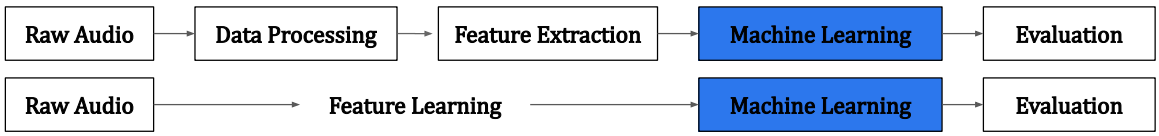



Figure 2.2: An overview of a feature-based machine learning pipeline (above) and an end-to-end deep learning pipeline (below), applicable but not exclusive to audio-driven tasks.

learning approaches often offer a more interpretable outcome as they have been derived from expert-driven domain knowledge, which may be more suited to sensitive targets such as emotional wellbeing analysis.

From left to right of the Figure 2.2, at first, the raw audio representation is seen as an input to the system. Ideally, with supervised approaches, the dataset from which the audio is taken will include extensive annotations in order for the machine learning model to learn representative patterns within the data. Such labels will be given directly to the machine learning algorithm unless a fusion of subjective labels is needed, or an imbalance in class distribution is likely to cause a bias.

Several tasks are performed during the data processing step. These include, but are not limited to, filtering e. g., noise removal, loudness normalisation, removal of silence, Voice Activity Detection (VAD), knowledge-based segmentation, and partitioning. For such data, fixed partitions are common, mostly as it is vital that partitions are speaker-independent and that no development is based on mixing speakers across the partitions due to the unique representation of individual voices. With that in mind, labelled meta-information may also be utilised for condition-specific partitioning strategies and ensure a balance of demographics such as sex and age (which are prominent in the voice [50]).

Once data is processed, numerous feature types can be extracted from the signal, where the particular audio target is essential to consider. For example, for environmental audio data, features derived from spectrogram images, or spectral derived acoustic Low-Level Descriptors (LLDs) such as Mel-Frequency Cepstral Coefficients (MFCCs) may be of use as they offer a more general overview of the acoustic environment. However, for speech, much of the acoustic environment should be ignored, and so LLDs which relate to speech characteristics, e. g., prosody and voice quality, may be more applicable. With this in mind, hand-crafted sets are also available, such as the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [51], which has been designed explicitly to capture emotional speech characteristics.

The next step in the pipeline is training the machine learning algorithm. In this thesis, primarily predictive models will be discussed, where the goal of the model is to learn a function $f$ which can determine the label value $y$ of an input $x$, as $y = f(x)$. This function

can be learnt in several ways, but two primary approaches are 'classification', or 'regression', where a classification algorithm attempts to learn categorical values, e. g., happy or sad, and a regression algorithm learns a dimensional value, often time-dependent in nature. However, the choice of model is not based on a static rule, and optimisation during the experimental development is vital to explore to find the best method for modelling and discussing the phenomena in question. For classification, in the computer audition field, researchers will apply both classical machine learning algorithms as well as deep learning networks. It may be that a less complex algorithm, e. g., Support Vector Machine (SVM) [52], is applied as a more interpretable tool for initial exploration where analysis of the phenomena is more of a priority. Deep learning, on the other hand, may be more suitable for larger-scale experiments, which are focused only on the accuracy to which a target can be recognised computationally, in this case, as well as the more straight-forward Feed-Forward Neural Network (F-FNN) the Convolutional Neural Network (CNN) has been successfully applied in the context of large-scale audio classification [53]. When it comes to regression tasks, the practice is similar, with SVM based regression algorithms being commonly applied at earlier stages due to their robust nature on smaller quantities of data. However, in this case, great strides are being made which are more specifically applicable to the audio signal, e. g., the Recurrent Neural Network (RNN), which can capture more fine-grained time-dependent relationships in audio [54], and results in this area continue to improve, particularly in relation to speech emotion recognition [55].

After a model is trained, it can then predict with a certain degree of accuracy values for unknown audio data. There are several ways to evaluate model performance. For classification, evaluation can be made with the 'accuracy' metric where a percentage is obtained based on the total correct instances overall observations. However, this can misrepresent the data, mainly if the classes are imbalanced. Therefore to consider the performance on a single class, the 'precision' and 'recall' can be calculated and combinations of this can give finer grained understanding i. e., the $F_1$-score, or the Unweighted Average Recall (UAR) (which is commonly applied to imbalanced speech datasets). For regression tasks, correlation (e. g., Pearson's correlation coefficient ($r$)) or error-based metrics (e. g., Mean Absolute Error (MAE)) can be applied to observe the relationship between two continuous streams of data (actual values vs predicted values). As with the model selection, the evaluation metric is also crucial to ensure that model performance is fairly reported – often, best practice, at least when tuning a model, is to observe multiple metrics simultaneously.

As mentioned above, a crucial part of the computer audition machine learning pipeline is appropriate data processing, and there are several computer audition sub-domains which are focusing only on data processing, as this can not only improve machine learning results but

also be applied as standalone models to improve user experience, e. g., speech enhancement technologies can, on the fly, improve the clarity of speech, primarily via background denoising [56] or packet loss concealment approaches [57]. Denoising is a commonly applied process in computer audition, as the complex nature of audio means that it is not uncommon to have disturbances such as traffic noise or recording artefacts. Concerning this, VAD is also applied to detect who is speaking and for how long. However, separating speakers on the fly remains a challenge, and for many years the computer audition community was tackling these problems via a combination of filters and classical algorithms e. g., hidden Markov models (HMMs) [58]; however recently end-to-end approaches show promising results [59]. In sterile settings, such as the office or home, these approaches are much more successful. However, where actual advancements continue to be needed is concerning in-the-wild data, and a number of machine learning challenges (i. e., competitions) are currently focusing on this as it pertains to emotion [60]. This is mainly a problem when analysing states of wellbeing, as truly spontaneous interactions are more accessible via in-the-wild data sources.

## 2.2 Computational Paralinguistics and Wellbeing

Computational paralinguistics is a sub-field of computer audition and speech processing in general, which applies computational approaches to the analysis and reconstruction of 'paralinguistic' phenomena. The field of paralinguistics is defined adjacent to 'linguistics' and does not necessarily include conventional linguistics, such as the structure of a language or phonetics [1]. Essentially, paralinguistics is concerned with *how* something is said and not with *what* is said.

This field began to gain broader attention as approaches for automatic prediction of affect in speech with machine learning become valid, and competitive challenges within the machine learning community began to focus on this topic. The first of its kind in this regard was the 2009 Computational Paralinguistics Challenge (ComParE) [61]. This challenge focused entirely on emotion data, and later iterations of ComParE included more extensive paralinguistic states and also traits [62]. In their book [1], Schuller and Batliner describe traits as being distinguishing qualities and inherited characteristics, essentially long-term aspects of an individual which are to an extent, fixed, e. g., age, or native language. However, a state is described as a short-term condition of being e. g., emotion or degree of stress. The authors also describe mid-term states, which are partly self-induced, and can be seen as traits for a period of time e. g., speaking style in a given social situation or intoxication.

Computational paralinguistics is particularly suited to the analysis of emotional wellbeing, due in part to the role which physiological and cognitive systems play during speech

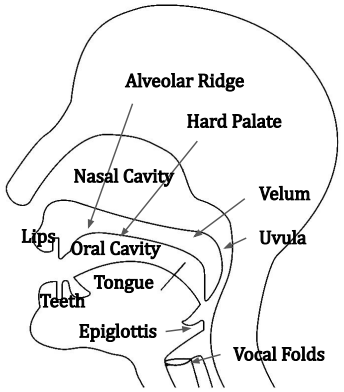



Figure 2.3: An illustration adapted based on [64], of the anatomical structures and organs involved in the mechanics of speech. The figure depicts the overall complexity of speech production, which supports the findings that specific conditions can alter vocal qualities by a substantial amount.

production – in other words, subtle changes in a speaker's mental state can subconsciously change the mechanics of our vocal apparatus, altering acoustic characteristics which can then be measured [7]. To detail this, in Figure 2.3 an anatomical representation of organs involved in speech production is given, and it is clear from the point of vocal production (the vocal folds) that there are several physical organs to which the signal should pass, and some will alter due to mental changes e. g., tension in the jaw during stress [63].

As already mentioned, the ComParE challenge is one such testing bed for exploration into these paralinguistic states and traits of speech, with a vast number of phenomena explored to a high degree over the years. Given the nature of ComParE, numerous machine learning approaches have been explored, with many more applying now deep learning-driven approaches when data quantities allow [7]. In Table 2.1, an overview of datasets which have been explored within the ComParE paradigm, concerning short to mid-term states of wellbeing, is given. As can be seen, many state-based tasks apply classification as either the absence or presence of a specific phenomenon. Although in many cases this may be suitable, finer-grained regression-based learning is becoming more of a standard for modelling states of speech [65], and a combination of the two is showing to be even more valuable when it comes to in-the-wild recognition [66].

The datasets listed are mainly containing a single native language, e. g., English or German. It is, of course, meaningful to maintain balance over such variables; however, particularly for traits, e. g., sex or states, e. g., of intoxication, language does not play a substantial role in recognising the phenomena. Although native language can be reasonably recognised with current computational methods [76], it is rather cultural groupings, which change how states of emotion are perceived in a vocalisation [77]. This may become more

Table 2.1: An overview of available datasets which target short to mid-term states of wellbeing and have a benchmarkable baseline provided by the Computational Paralinguistics Challenge (ComParE). Table includes, detail of the sub-challenge task itself, spoken-language, number of participants (#), gender (f)emale, and mean ($\mu$) and standard deviation ($\sigma$) of age in years.

| **Dataset** | **Year** | **Task** | **Demographics** | **Language** |
|---|---|---|---|---|
| SLC | 2011 [67] | 2-class, presence vs absence of fatigue. | # 99 (56 f), age $\mu$ 24.9 ($\sigma$ 4.2) years. | German |
| ALC [68] | 2011 [67] | 2-class, presence vs absence of intoxication. | # 162 (78 f), age $\mu$ 31 ($\sigma$ 9.5) years. | German |
| GEMEP [69]. | 2013 [70] | 12-class, emotion see [70]. | # 10 actors (5 f), –, –. | German |
| MBC [71] | 2014 [72] | 2-class, low vs high physical load. | # 19, (4 f), –, –. | German |
| CSLE [73] | 2014 [72] | 3-class, 1,2,3 degree of cognitive load. | # 26 (6 f), –, –. | English |
| URTI [74] | 2017 [75] | 2-class, presence vs absence of cold symptoms. | # 630, (248 f), age $\mu$ 29.5 ($\sigma$ 12.1) years. | German |
| USoMS | 2018 [29] | 3-class, low, mid. and high valence. | # 100 (85 f), age $\mu$ 22.3 ($\sigma$ 3.6) years. | German |
| SLEEP | 2019 [30] | Karolinska Sleepiness Scale (KSS) 1-9 see [30] | # 915 (364 f), age $\mu$ 27.6 ($\sigma$ 11) years. | German |
| USOMS-e | 2020 [9] | 3-class, low, mid. and high arousal or valence. | # 87 (55 f) age $\mu$ 71.01 ($\sigma$ 9.14) years. | German |
| UCL-SBM | 2020 [9] | Continuous breath rate. | # 49 (29 f), age $\mu$ 24 ($\sigma$ 10) years. | English |

complex when considering para-language i. e., laughter, screaming or crying, which can express a variety of emotions and meaning [78] but acoustically appear similar.

### 2.2.1 A Baseline for Computational Paralinguistic

In most cases, the baseline for speech-based challenges, such as the one provided in the context of Computational Paralinguistics Challenge (ComParE), would be set based on extraction of hand-crafted features, and a supervised (labels are provided) machine-learning algorithm e. g., an SVM (see Chapter 3). This method is a particularly suitable set-up for smaller sized datasets, partially as meaningful and robust insights can be obtained quickly. However end-to-end approaches, e. g., utility of the End2You toolkit [79][1] or similar, are becoming more common, and baselines have been release with end-to-end methods since

[1] github.com/end2you/end2you, accessed on: 09.2021

the 2018 ComParE challenge [29]. End-to-end methods are shown to be meaningful for continuous tasks, applying a CNN to extract spatial features, and an RNN-block architecture to model the time-dependencies of the audio data [79]. However, given the limited data available, robust performance can be difficult to obtain, and optimisation can be slower compared to traditional machine learning methods, never-mind the computational expense which can be limiting for some academic groups.

In regards to available features, there are several hand-crafted feature sets which have become standard for paralinguistics recognition, and remain a solid approach. Such feature sets are derived from LLDs which were selected in collaboration with experts, particularly those who have a phonetics or linguistics background. A very prominent hand-crafted feature set is the ComParE set [80], which is known as a brute-force approaches as it consists of 6 373 functionals, which are the result of computations over 65 LLDs contours, and has been successfully applied in a wide range of paralinguistic tasks. Other sets, as mentioned earlier, include eGeMAPS [51], and the openSMILE toolkit can be used to extract these configurations and more and is still popularly used [2]. openXBOW [81] is another popular feature extraction toolkit [3], which generates a bag-of-words representation from acoustic LLDs, and often within the context of ComParE, this feature quantisation method has achieved the strongest baseline result, likley as it retains only the most meaningful information. It is also worth noting that similar to openSMILE, is the python native audio processing toolkit Librosa[4], which can be applied for extraction of many acoustic LLDs, and much more, e. g., audio plotting and audio conversion.

Deep learning approaches are also being applied to extract features from paralinguistic phenomena with some great success [82]. One prominent feature set popularly applied for speech-based tasks, and now paralinguistics is VGGish [53], which is an audio adaptation of the vision-based VGG16 with some layer pruning to reduce the number of parameters. For VGGish was initially presented as the embeddings for the large-scale AudioSet [83] dataset, and the network was initially pre-trained on audio from the YouTube 8M dataset [84], making this set suited to environmental audio use cases, and it has recently been found to be suitable for Speech Emotion Recognition (SER) tasks [60]. Similarly to the VGGish approach, several researchers have also been utilising image-based pre-trained CNNs to extract activations from audio plots, e. g., spectrograms. One popular toolkit available for this is DeepSpectrum [85], which essentially feeds spectrograms images to a pre-trained CNN, e. g., AlexNet, and then extracts features from the fully-connected layers of the CNN. Further details on these approaches will be given in Chapter 3.

[2] github.com/audeering/opensmile, accessed on: 09.2021
[3] github.com/openxbow/openxbow, accessed on: 09.2021
[4] github.com/librosa/librosa, accessed on: 09.2021

When discussing computational paralinguistics baselines, it is essential also to note that in recent years, many approaches are derived from multimodal data, often with a substantial improvement found for specific targets. This is particularly true when it comes to emotional targets, where linguistic features have been found to model positivity (i. e., emotional valence) better than the audio signal. In ComParE 2020 for the first time, a language transcription feature extractors was provided with the baseline and reported a substantial improvement over audio-only for classification of mood in elderly speech [9]. Furthermore, through fusion with video-based features, paralinguistic targets are finding marginal improvement in the area of stress [15]. These multimodal approaches do all show great promise, however, such methods tend to lean away from pure-paralinguistic research, and become more brute-force in nature, leading to less interpretation.

## 2.3 Emotional Speech and Empathy in AI

This thesis is concerned with computational audio-based methods for analysing and understanding emotional wellbeing, and although it is not the core focus, it would be a natural fit to discuss that a use case for this research is to support development into the empathy of AI systems. Empathy, has been discussed in the AI related literature for quite some time, particularly as it pertain to the adaption of pedagogical agents [86], however as AI become more embedded in daily life the discussion of empathy is becoming more prominent [87, 88]. Empathy is often included as part of the next evolution in AI, Artificial General Intelligence (AGI) [89] – i. e., the ability for an AI to perform human intellectual activities. Along with being able to perceive, access, and decipher reality, AGI, should be emotionally aware and consider 'emotional grading', i. e., the emotionality of its given 'reality', and behave with empathy towards emotion as needed. This does not only pertain to human interactions with physical AIs, e. g., robotics; this can also mean that the AI system within a smart-device for example, is more considerate of a given human state and designed sensitively, e. g., understanding general digital wellbeing. In this way, empathy and AI is informed and slightly adjacent to the well-established area of affective computing [90], which focuses specifically on intelligent methods to interpret and understand affective states. The empathy of AI in a way extends on this, in that it is aimed at developing more harmonious interactions, informed by researchers from both technical and non-technical background, e. g., Googles Empathy Lab. AI empathy includes more than states of mood and feeling, but also current activity, interpersonal relationships, and objective truths e. g., how full an individuals calendar might be that day. Such aspects may be having an impact on an individual's wellbeing at a given

time, and therefore responding with empathy may be a better long-term solution for our interactions with AI technology.

The criticism for empathy and AI should also be noted here, mainly as it concerns healthcare. Several experts in the health domain consider that empathic human behaviours are both unethical and impossible to replicate [91]. With this in mind, it is essential to define the scope of empathy in AI within the context of this thesis, and when considering human emotional wellbeing, the benefit for this as well as the true technical meaning should be clear. One area highlighted for concern relates to the definition of empathy itself, for which there are three core descriptions, affective (emotional) empathy, cognitive empathy, or motivational empathy. Emotional and motivational empathy are types of empathy that elicit a visceral feeling. In contrast, cognitive empathy can be somewhat more of a means to an end, acting on what is seen and not based on emotional understanding; as the authors in [91] mention, this can be a manipulative trait, and therefore avoiding this type of empathy which may respond to a reward cycle could be critical. In general, however, in the context of this thesis, when discussing empathy and AI this is not focused on applications that are for human replacement, but instead resources for human support, which of course can still be questioned for it ethical implementation, however this is why interdisciplinary with the field is crucial, as it pertains to the RQ-4 (see Chapter 1), and later discussed again in Section 5.2. In the sense of aiding and not replacing, by understanding a state of emotional wellbeing from audio, this information can be a tool for improving empathic responses, which in turn aids individuals during the challenges of modern life, and therefore facilitates agency for improved emotional wellbeing – in other words, when considering the earlier definition of wellbeing detailed in Figure 1.1, suppose that an empathic AI state is an additional resource which is available during a short-term increase in challenges, restoring the balance of wellbeing.

### 2.3.1 Modelling Emotion from Speech

With the above passages in mind, understanding emotion is a core component of an AIs ability to reason with emotional empathy, and Speech Emotion Recognition (SER) as a research area would be a vital aspect of this. SER is an adjacent area of machine learning to computational paralinguistics, which targets either categorical (i. e., discrete), or continuous states of emotion from the speech signal. This section is an extension to many of the points already discussed in Section 2.2, with a focus on, and introduction to speech emotion modelling. In Figure 2.4, a step-by-step outline for a SER pipeline is shown.

As a first step when targeting emotion, a model for emotion should be found; this is typically a selection categorical emotion classes or a dimensional representation of an emotions' activation. There are a number of models, which are continually discussed in-depth

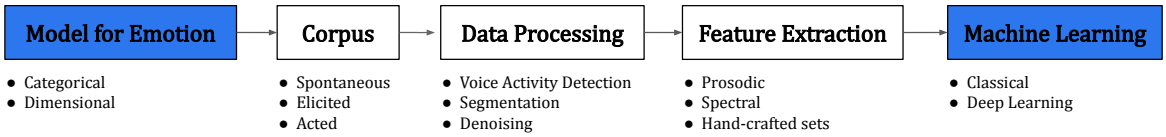



Figure 2.4: Components for the development of an SER system. Adapted based on the overview presented in [92].

for their relevance by the emotion community [93], however for the purpose of this thesis, those which are currently being regularly applied within SER and affective computing in general will be discussed in the proceeding section. Furthermore, the methods for rating an utterance are also continually discussed, i. e., self-assessment of one's emotion vs the perceived emotion from others, and there is debate on the validity of models based on perceived emotions, particularly as ones true feelings can often be masked [94]. However, given that perception is essentially what a machine is attempting to imitate several researchers continue with this line of thinking and utilise the perceived emotion ratings, yet many datasets (including several described in this work) will attempt to include various ratings for which a model can be tuned too.

As it pertains to categorical emotions, Ekman's, so-called 'basic six emotions' [95], continue to be used and includes the emotions of, anger, disgust, fear, joy, sadness, surprise. Such emotion models from a machine learning perspective result in quite course outcomes which may only be applicable for specific use cases. To overcome this affective computing researchers began to apply dimensional models such as the Russel's Circumplex model of affect [96], which considers the activation (arousal) of the emotion on the vertical axis, and the positivity (valence) of the emotion on the horizontal axis (see Figure 2.5). This model allows for a fine-grained approach to emotion modelling, which might be more intuitive for time dependent audio and the fluctuations in emotion that occur during speech. Russell's model has been widely adopted in SER, with speech having particular success at modelling arousal [1]. However, this model for emotion is widely discussed for its validity [97] with researchers in emotion continuing to develop various other models such as Cambria's Hour Glass of Emotions [98], a biologically-inspired and psychologically-motivated model which incorporates ideas from established discrete and dimensional models.

Once a model for emotion is selected, a dataset should be acquired (or collected) with these ratings associated to the data. In Table 2.2, an overview of a selection of datasets that can be used for SER is given. Typically, the speech within such datasets would be gathered in three ways, i) spontaneously (or in-the-wild), ii) via an *elicitation* method iii) acted . Most researchers consider that spontaneous vocalisations which have been gathered in-the-wild are a better representation of true emotion. However, spontaneous recordings are challenging to

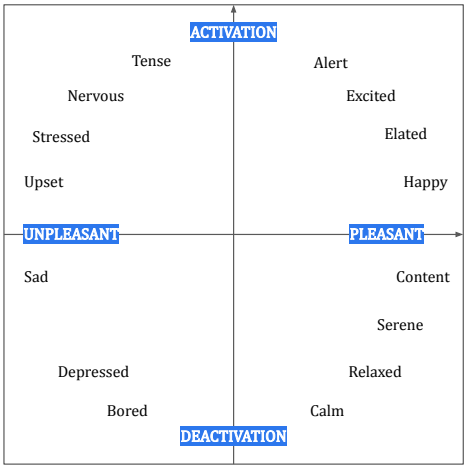



Figure 2.5: An illustration of Russel's Circumplex model for Affect, adapted based on the model described in [96]. In this figure, the emotional valence (e. g., positivity) is represented along the horizontal axis, and the emotional arousal (e. g., activation) is along the vertical.

Table 2.2: A selection of SER dataset for academic research. Including, number (#) of speakers, emotion type (Spont)aneous or (Elic)ited, the (Emo)tion model as (Dim)ensional or (Cat)ergorical, the (Modal)ities available, (A)udio, (V)ideo, (T)ext, and the spoken-languages.

| **Dataset** | **#** | **Type** | **Emo** | **Modal** | **Language** |
|---|---|---|---|---|---|
| MuSe-CaR [32] | 70 | Spont. | Dim. | A, V, T | English |
| MSP-Podcast [99] | 100 | Spont. | Dim., Cat. | A | Multi |
| SEWA [100] | 398 | Spont. | Dim. | A, V | Multi |
| DEMoS [101] | 68 | Elic. | Cat. | A | Italian |
| RAVDESS [102] | 24 | Acted | Cat. | A, V | English |
| CMU-MOSEI [103] | 1000 | Spont. | Cat. | A, V | English |
| RECOLA [104] | 46 | Elic. | Dim. | A, V | French |
| GEMEP [69] | 10 | Acted | Cat. | A | French |
| IEMOCAP [105] | 10 | Acted | Cat. | A, V | English |
| EMO-DB [106] | 10 | Acted | Cat. | A | German |

obtain; even with the advent of social media, which offers huge quantities of video and audio data, annotating data remains a time-consuming effort. Spontaneous data may also come with additional data processing requirements, such as denoising or source separation. On the other hand, elicited emotional datasets are typically gathered in a lab setting, with participants given tasks to perform, such as reading a book or observing images that are designed to elicit a particular emotion. This style of speech collection is often criticised for being somewhat mild in nature due to the ethical limitations of researchers provoking strong negative emotions. Actors, however, are much more versed with a wide variety of emotions, and particularly in earlier stages of SER and computational paralinguistic research, such datasets allowed for

very stable and interesting insights into the manifestation of emotion in speech [106]. With time, the field has moved away from acted speech to more spontaneous, in-the-wild sources. Acted speech may be less representative of true emotion, and particularly when considering empathy and AI this may bias the model towards strong emotional responses, which are less likely to occur in daily-life.

Despite this demand for more spontaneous, and natural speech, in regard to academic purposes at least, the available data is limited. This is due to a number of the factors already mentioned, but primarily the time expense that is associated to audio data collection and annotation. One method which can be applied to capitalise on the already gathered smaller datasets, are transfer learning methods, which learn and combine multi-domain representations of multiple datasets. In [107], the authors do just this, by combining a total of 26 SER datasets they propose EmoNET, which is based on a vision derived approach, where a combination of a deep ResNet architecture and residual-adaptors [108], are used to learn representations. Within this work [108], the authors combine the various emotion-model classes within each of the individual datasets by mapping them to the classes derived from the dimensional models of emotion (arousal and valence), i. e., low or high arousal, and negative or positive valence, and neutral. Results obtained via this approach are consistently above chance-level for each of the paradigms explored, showing that this may be a promising strategy to allow for broader representation to be learnt, however language and cultural dependencies as well as differences in the perception of the raters of each dataset, which would be fruitful next steps.

For recognition of emotional states in general the state-of-the-art is continuing to evolve, but primarily the focus in recent years has been inclusion of deep learning, particularly attention-based Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) models [109]. In this regard, in [109], the authors propose a dual-attention based Bidirectional Long Short-Term Memory (B-LSTM), with mel-spectrogram as input, and obtain a strong accuracy up to 2% over next best approach from [110] (also an attention-based LSTM-RNN architecture), on the well-known IEMOCAP dataset. However, that is not to say that meaningful results can not be obtained utilising more conventional machine learning approaches, e. g., the SVM, and for wellbeing-based application where data is even more scarce than general speech emotion datasets such algorithms may allow for more interpretability, which may be needed given a specific task focus.

# Methodology

In the proceeding chapter, details are given for the fundamental aspects of methodologies applied in the experiments of Chapter 4. The order of sections follows the typical machine learning pipeline from acquiring data to evaluating the model performance, see Figure 3.1. As with Chapter 2 a description will be given at the start of each section for the term itself.

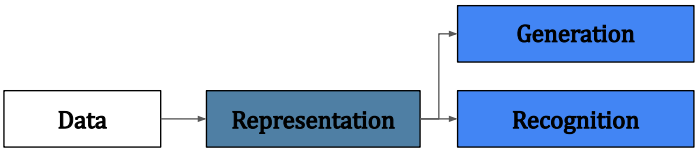



Figure 3.1: An illustrated overview of the outline for the proceeding sections as it relates to the components of machine learning pipeline.

## 3.1 Data

Acquiring data is a crucial aspect for any machine learning and Artificial Intelligence (AI) system, and in relation to modelling human states, this can include textual, video (or image), and in the case of this thesis, audio of any human phenomena. From a philosophical perspective, data can be considered to be assumed facts which are the basis of future reasoning, and so when applying this to machine learning the quality of data – in terms of how well does it represent the population or how noisy is it - is an important aspect to consider. In the following, first, a selection of datasets that have been collected in the context of this thesis will be detailed, followed by a description of methods for annotation as they relate to the experiments of this thesis, and then as data scarcity is a common bottleneck in the domain of computer audition, computational paralinguistics, and particularly emotional wellbeing from speech, a description of popular data augmentation methods is given.

### 3.1.1 Acquired and Applied Data

Throughout the development of this thesis, numerous speech and multimodal datasets have been sourced and collected within the theme of emotional wellbeing. In this section, a brief

Table 3.1: An overview of the datasets gathered during the time of this doctoral research. Detail is given for, type (spont)aneous or elicited, number of participants (#), mean ($\mu$) and standard deviation ($\sigma$) for age in years, labels (targets), (Modal)ities available, (A)udio, (V)ideo, (T)ext, or (Ph)ysiological signals, and spoken-(lang)uage. * indicates a focus on states of emotional wellbeing, and † indicates its utility for experiments in Chapter 4.

| Corpus | Type | Demographics | Targets | Modal | Lang. |
|---|---|---|---|---|---|
| SinS-DB [28] | Acted | # 32 (17 f), age $\mu$ 29.8 ($\sigma$ 9.9) years. | Perceived sincerity | A | English |
| BioS-DB [23] ∗† | Spont. | # 55 (22 f), age $\mu$ 28.9 ($\sigma$ 10.5) years. | Continuous arousal and valence. | A, Ph | English, German |
| USoMS [29] ∗ | Spont. | # 100 (85 f), age $\mu$ 22.3, ($\sigma$ 3.6) years | Self-assessed, categorical arousal. | A | German |
| USOMS-e [9] ∗ | Spont. | # 87 (55 f), age $\mu$ 71.01 ($\sigma$ 9.14) years. | Self-assessed, categorical arousal and valence. | A,T | German |
| FAU-TSST [23] ∗† | Elicited | # 43 (29 f), age $\mu$ 24.26 ($\sigma$ 4.97) years. | Sequential saliva-based cortisol samples. | A,Ph | German |
| Reg-TSST [31] ∗† | Elicited | # 27 (14 f), age $\mu$ 22.74 ($\sigma$ 2.96) years. | Sequential saliva-based cortisol samples, and continuous heart rate. | A, Ph | German |
| Ulm-TSST [60] ∗† | Elicted | # 69 (49 f), age $\mu$ 25.06 ($\sigma$ 4.48) years. | Sequential saliva-based cortisol samples, continuous emotion (arousal, and valence), and continuous heart rate, respiration rate, and Electrodermal Activity (EDA). | A, V, L, Ph | German |
| MuSe-CaR [60] | Spont. | # 350 reviews from 70 host speakers | Continuous arousal, valence, and trustworthiness. | A,V,T | English |
| DAC [19] ∗† | Spont. | # 252 (48 f), age $\mu$ 31.5 ($\sigma$ 12.3) years. | Beck Anxiety Inventory (BAI) | A | German |

overview of each dataset that is used in the proceeding chapters will be given. As a more general overview, in Table 3.1 all datasets collected during the time of the thesis are detailed. It is also important to note that all the datasets that have been used within experiments described in this thesis received ethical approval from their respective universities, and all participants gave informed consent.

#### 3.1.1.1 Trier Social Stress Test Data

During this doctoral thesis, three German-speaking datasets were gathered under the renowned Trier Social Stress Test (TSST) [111], namely, the Friedrich-Alexander-Universität-Trier Social Stress Test (FAU-TSST), the Regensburg University-Trier Social Stress Test (Reg-TSST), and the Ulm University-Trier Social Stress Test (Ulm-TSST). In order to avoid

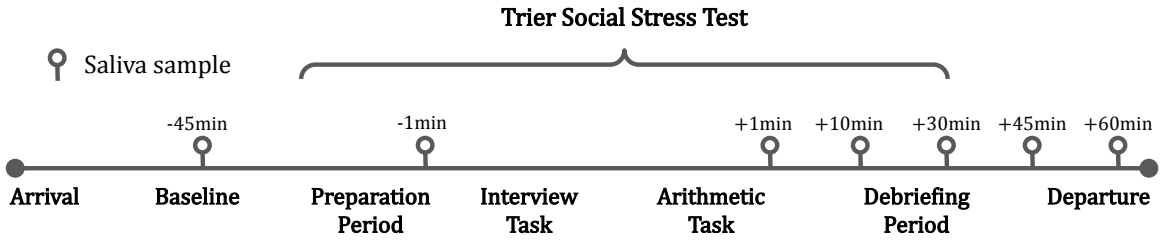



Figure 3.2: An overview of the typical TSST paradigm, applied in FAU-TSST, and Reg-TSST, where Ulm-TSST excludes the arithmetic task.

repetition, the general procedure for the TSST study will be described once, and where needed, general details about each of the three datasets will be highlighted. The experiments performed for these datasets will be detailed in Chapter 4 Section 4.1, based on previously published works, where the FAU-TSST dataset was first introduced in [112], and the Ulm-TSST dataset in [15], with the Reg-TSST first introduced in [31]. In [31], for the first time the three datasets were combined for cross-corpus analysis.

The TSST is a renowned and highly standardised test used (amongst other things) for the analysis of emotional stress [113]. The test at its core includes a speaking task and typically several physiological signals e. g., saliva-based cortisol samples, will also be captured, making it a beneficial testing paradigm in the context of computational analysis of states of emotional wellbeing.

In Figure 3.2 an overview is provided for the typical TSST testing paradigm. Before arrival at the TSST site, instructions for the subjects included refraining from exercise, eating, and drinking anything except water before arrival. Verbal and written instructions are given to the subject when they arrive at the test site, followed by a resting period. During this time, for the FAU-TSST and Reg-TSST a saliva sample (S0 45 minutes before TSST) was collected as the participant's cortisol baseline. The subjects are then guided to the test room and introduced to observers wearing white lab coats, and instructed to prepare to speak for five minutes, presenting themselves as the best candidate for a vacant job. One minute before the interview task begins, another saliva sample is taken (S1 -1minute), and recording begins when they enter the room. After this, for a further 5minutes in the FAU-TSST and Reg-TSST datasets, subjects are given a mental arithmetic task, where they should serially subtract 17 from 2 043. After completing the TSST speaking tasks, six more saliva-based samples are taken from the subjects (S2-S7).

The saliva-based cortisol samples are measured in nanomoles per litre (nmol/L). The immunoassay (i. e., biochemical analysis procedure) applied to extract cortisol from the saliva samples varied, where FAU-TSST utilised the Chemiluminescence Immunoassay

(CLIA), and Reg-TSST the Dissociation-Enhanced Lanthanide Fluorescence Immunoassay (DELFIA). This means that the derived cortisol values are not entirely comparable, for further detail on the difference in these procedures, the interested reader is directed to [114]. For the Ulm-TSST and Reg-TSST datasets, there is also continuous physiological signals available, heart rate as Beats per Minute (BPM) from both, and for the Ulm-TSST dataset only, there is EDA, and respiration (based on chest displacement). Further to this, the Ulm-TSST dataset includes continuous emotion ratings, which were rated by three annotators for the dimensions of arousal and valence, at a 2 Hz sampling rate.

#### 3.1.1.2 Düsseldorf Anxiety Corpus

The Düsseldorf Anxiety Corpus (DAC) was collected by members of the Institute of Experimental Psychophysiology, Düsseldorf, Germany, and was first introduced in [19], and in Chapter 4 Section 4.2, these experiments will be detailed. The dataset includes 252 speakers aged 18 to 68 years old (average of 31.5 years, standard deviation of 12.3 years) performing various vocal exercises. The files are categorised into different phonations, including sustained vowels, read, and free speech. 239 of the speakers in DAC were evaluated under the Beck Anxiety Inventory (BAI) questionnaire [115]. During the BAI, individuals answer a series of questions relating to their wellbeing in the last 30 days, on a scale from 0–3. A total of under 21 indicates anxiety, and above 36 indicates potentially concerning anxiety levels.

Table 3.2: Overview of the DAC, including gender distribution, BAI class (Low, High), feeling of choking (No) or (Has) symptoms, difficulty in breathing (No) or (Has) symptoms.

| | **BAI** | **Breathing** | **Choking** |
|---|---|---|---|
| *Male:Female* | *Low: High* | *No: Has* | *No: Has* |
| 69:170 | 191:43 | 193:46 | 209:30 |

In relation to speech, there are two questions within the BAI which specifically relate or may effect vocalisation, e. g., "Have you experienced feelings of choking?" and "Have you experienced a difficulty in breathing?". With this in mind, groupings in the data have been prepared, for subject which do have and do not have such symptoms, as well as those above (high) and below (low) the 21 point threshold for anxiety presence (see Table 3.2 for detail).

#### 3.1.1.3 The BioSpeech Database

The BioSpeech Database (BioS-DB) was first introduced in [23], and later updated in [25], and the experiments of [25], will be described in Chapter 4 Section 4.3. The BioS-DB was

collected at the University of Augsburg, Germany and obtained full ethical approval [1], and the latest version of the data is available for research by restricted access [2]. Essentially, the initial aim of the BioS-DB was to explore physiological and speech signals during states of performance anxiety, which occur when speaking aloud in front of others. Version 2, will be utilised within this thesis and includes 42 participants (17 female), with a $\mu$ age of 26.76 years ($\sigma$ of 6.62 years), speaking publicly in front of a group of observers (minimum 4). Of all of the participants, 30 of the 42 are native German, and 12 are from foreign countries.

Each participant was asked to speak the text ("The North Wind and the Sun") aloud in German and then in English. During their speech, three of the observers in the room were using joysticks to continually rate the emotion of the speakers (2 Hz sampling rate), where the vertical axis of the joystick relates to the emotion activation (arousal), and the horizontal axis is the positivity or valence of the emotion. The raters were students who had undergone an introduction training session on core concepts of the dimensions of arousal and valence. The original gold-standard for the ratings was calculated with Evaluator Weighted Estimator (EWE) (see Section 3.1.3.1) and the mean inter-rater agreement across all speakers in the BioS-DB, was .47 and .36 arousal and valence respectively, (based on a range of [0,1]). In Figure 3.3, the distribution of the gold-standard ratings for both emotional dimensions, across all speakers used in our experiments is given.

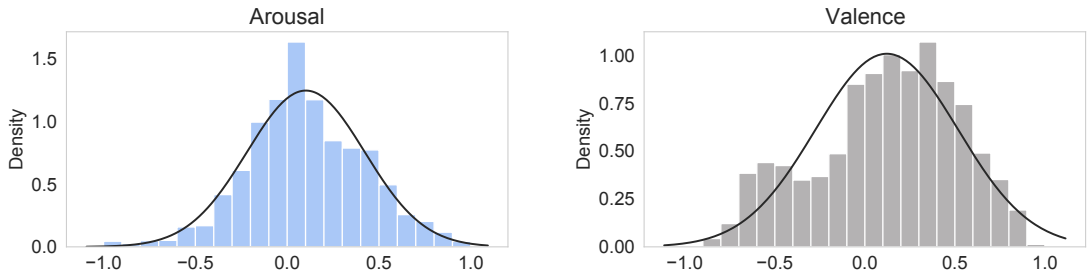



Figure 3.3: The distribution (and normalised distribution curve) of the gold-standard ratings for arousal (left) and valence (right) across all speakers in the BioS-DB.

During their speech, two audio channels were captured, one from a lapel microphone (AKG C417 L) and one from a room microphone (C562 BL) placed on the table in front of the speaker. Furthermore, two sensors were placed on the fingers of the participants to capture Blood Volume Pulse (BVP) as a % of blood volume pressure, and Skin Conductance (SC) measured in microSiemens ($\mu$S), at a sampling rate of 2 048 Hz and 256 Hz, respectively.

[1] Ethical approval obtained in 2018 from the University of Augsburg's ethics commission under the project title 'Multimodal Signal Recording Techniques and Emotional Analysis'.

[2] 10.5281/zenodo.4281253 access on: 09.2021

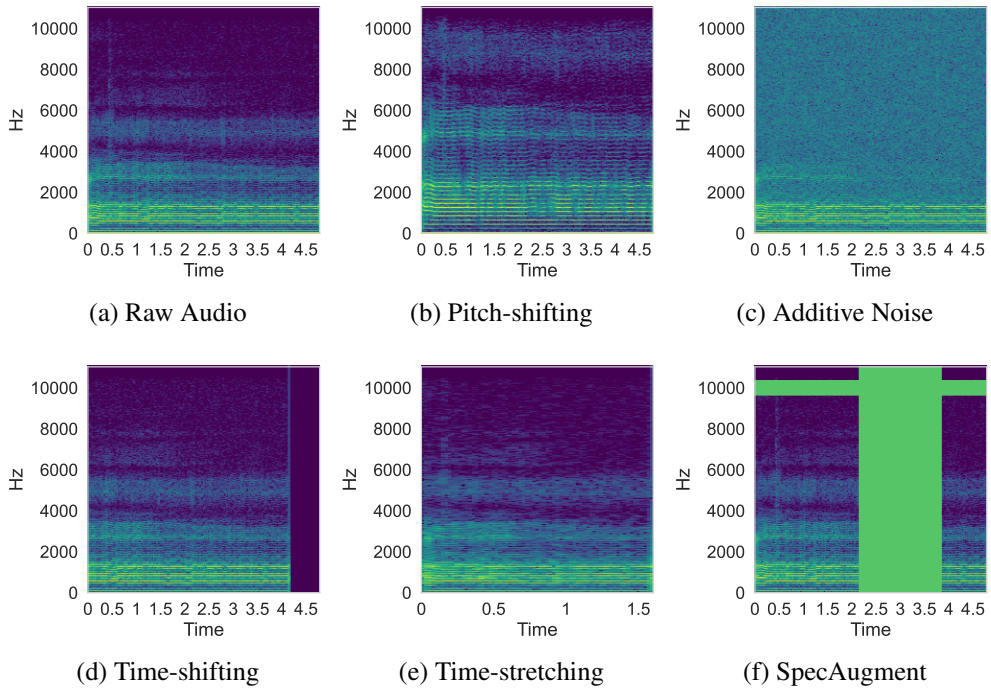
#### 3.1.1.4 The Geneva Multimodal Emotion Portrayals Corpus

Of all the datasets described here, the was not gathered or processed by the authors but has been utilised with the experiements of Chapter 4. Geneva Multimodal Emotion Portrayals Corpus (GEMEP) [69] was first utilised in the 2013 Computational Paralinguistics Challenge (ComParE) [70], and includes ten native French actors (five female) speaking nonsense utterances to avoid cultural, and lexical bias. Given well established use for emotion recognition tasks, this dataset was used for in a previous publication by the author [26], and the experiments for this will be described in Chapter 4 Section 4.4. For those experiments a subset was created, and includes only four of the available emotional classes, Hot-Anger (referred to as Anger), Elation, Sadness, and Pleasure. Those emotions were selected to cover the four quadrants of Russel's circumplex for affect [116] (e. g., Elation = High Arousal, High Valence, and Sadness = Low Arousal, Low Valence), offering a more distinct emotional setting, with potentially more perceivable diversity in the classes. The rating of the GEMEP dataset were obtained from 57 participants who rated the media from a combination of audio and/or video perception, and in [69] this explained in more detail.

### 3.1.2 Data Augmentation

Data augmentation methods are mainly focused on increasing the overall quantity of data available. This is particularly meaningful for audio, given its time-dependent nature, which means acquiring such data manually is time-consuming and costly, typically resulting in smaller scale datasets than other areas of machine learning research e. g., vision and text, which can more effectively be scrapped from in-the-wild sources e. g., the internet, and social media. There is however the AudioSet database, which has been collected in this manner, and is one audio-specific resource targeted primarily at acoustic even detection tasks, given the categorical labelling provided [83]. Despite the major benefits of audio, the sparsity of data when it comes to emotional wellbeing targets, means that many deep learning methods are not suitable, and robust results are a challenge to obtain on the raw data alone. Furthermore, imbalance across classes can be common, particularly in regards to demographics, e. g., male vs female, where students are the primary target. As well as this certain classes can be in a minority e. g., highly emotional states, which are more of a challenge to obtain.

As well as increasing overall size of a dataset, when classes within a dataset are imbalanced, several conventional approaches can be applied to the training set independent of the modality to either upsample the minority class or downsample the majority class, these include: i) *Random re-sampling*, a standard and easily implemented method, where the data points from the minority class are duplicated (upsample) or from the majority, class are




(a) Raw Audio (b) Pitch-shifting (c) Additive Noise

(d) Time-shifting (e) Time-stretching (f) SpecAugment

Figure 3.4: Example spectrogram representations of audio-based augmentation. a) The original source audio, b) pitch shifted by +10 Hz, c) SNR of 10dB, d) time-shifted by 0.2 seconds, e) time-stretched by a factor of 3, f) SpecAugment randomly generated sample.

removed (downsample). ii) *Synthetic Minority Oversampling Technique (SMOTE)*, a more targeted approach for upsampling [117] which artificially produces instances of the minority class based on the clusters which exist. SMOTE is selecting class data points and calculating the k-Nearest Neighbours (kNN) for this point, and synthetically adding points between the neighbouring classes. iii) *Tomek links*, a method to mitigate the inherent randomness from the other approaches which may not reflect the original distribution, which searches and removes pairs of data points based on their euclidean distance from one another.

#### 3.1.2.1 Audio-Specific Data Augmentation

There are several common methods for data augmentation which are specific to audio (see Figure 3.4), these include (but are not limited too):

*Pitch shifting* is the process of altering the pitch of the sound without altering the speed of the sound and without consideration to the fundamental pitch ($F_0$) of the voice. This should be applied sparingly in the context of speech as it may be destructive.

*Additive noise*, is typically the addition of white noise (e. g., random frequencies with regular distribution at the same volume intensity), is added to the signal at a set loudness. In a

similar way real-life 'background' noise, e. g., from an office environment, could also be added to the signal at a given Signal-to-Noise Ratio (SNR).

*Time-shifting* is the process of moving the signal by a set amount either forwards or backwards in time. Of all those described herein, this is typically the least destructive, and most similar to the original sample.

*Time-stretching* is the process of speeding up or slowing down the signal,without affecting the pitch of sound but altering the duration, this process can destructive in a similar way to pitch shifting, as it may alter natural aspects of the speech phenomena.

When applying manipulation to the audio signal as an augmentation approach, the choice should be based on the audio phenomena being targeted, as it will impact the machine learning models' ability to learn the nature of the embedding space in regards to a given phenomena. For example, applying a degree of pitch shifting when utilising the speech signal can alter the perceived state, particularly when it comes to emotion. Time-shifting is also a minimal alteration and may, in some cases, not improve the results, as no diversity is added to the embedding space, where more diversity is needed, additive noise may be more suited as this is subjectively very different to the source data [26].

When utilising spectrogram images for training, there is also the SpecAugment approach [118], which applies masking to the spectrogram images on the frequency and time-axis at random positions, as well as warping the image horizontally or vertically. This method has shown to be promising for several audio tasks, including for SER [25].

### 3.1.3 Gold Standards for Emotion

As discussed in Chapter 2 obtaining label information for computational paralinguistic and Speech Emotion Recognition (SER) tasks is a time-consuming task which requires in most cases a substantial amount of human-effort for reliable results. Within the above datasets, numerous labelling strategies have be applied in the case of subjective data. These are either categorical (e. g., the GEMEP) or continuous (e. g., BioS-DB), and for perceived emotion it is vital to consider the opinion of at the very least three external raters, depending on the task and the level of experience [119]. In some cases, self-assessed emotional values can also be utilised, however in the psychology literature, some findings suggest that self-reported dimensions of arousal do not correlate highly with physiological responses [120], and therefore are a challenge for audio [29].

In regards to subjective targets, a gold standard is the agreed upon signal and can be considered as a pseudo-ground-truth. When creating a gold standard, the aim is to build a consensus from several individual annotations. The methods for this is different for

categorical or continuous ratings, with some methods developed with subjective labelling in mind and other more standardised time-series-based approaches applied successfully to subjective-based tasks. The process can be relatively straightforward for categorical labels, and it is common to use a majority-voting approach, where the most commonly occurring values are considered, mainly when the values are not ordinal. For an ordinal discrete value, it may be more meaningful to calculate the mean across all raters. Various clustering methods can also be applied for discrete gold standard creation, and an approach for this was described as part of the 2021 Multimodal Sentiment Analysis in Real-life Media Challenge (MuSe) baseline [15], and made available with the release of the MuSe-Toolbox [27]. However, the effectiveness of any approach will depend on the number of and agreement between raters. Agreement (referred to as inter-rater reliability) is, therefore, a meaningful value to report, and in the case of multiple raters, one metric for categorical labels is Fleiss' Kappa [121].

For a dimensional (continuous) rating, there are several methods to create a gold standard. Some include weighting based on the agreement, and others provide compensation based on an annotator delay [122]. To calculate agreement between continuous ratings typically, researchers will apply a correlation metric such as Pearsons Correlation, however as emotion is highly subjective, considering the scaling variance between ratings is meaningful, and so the Concordance Correlation Coefficient (CCC) can also be reported. In general, the methods for gold standard creation are slightly more complex for continuous rating, and so methods applied to data utilised in the experiments of Chapter 4 will be described herein with detail.

#### 3.1.3.1 Evaluator Weighted Estimator

The Evaluator Weighted Estimator (EWE) [123] is one method which is very applicable for subjective rating as it is based on a reliability evaluation of the raters, as presented in [36]. EWE has been utilised for the ratings of the BioS-DB [25], and the Ulm-TSST [24], as well as the initial release of MuSe-CaR [32]. EWE essentially is a weighted mean of all rater-dependent annotations, sometimes interpreted as the weighted mean of raters' similarity [123]. To compute the weights, the cross-correlations of an annotation to the mean of all other annotations is calculated for each annotation. It can be formally expressed by:

$$\hat{x}_n^{\mathrm{EWE}} = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} r_k \hat{x}_{n,k}, \tag{3.1}$$

and then $r_k$ is the similarity of the $k$-th annotator to all other continuous ratings.

#### 3.1.3.2 Rater-Aligned Annotation Weighting

To extend on the benefits of EWE, in the context of emotion, the Rater-Aligned Annotation Weighting (RAAW) approach was first proposed by the author and colleagues in [15], and is included as part of the MuSe-Toolbox [27], which provides a number continuous and discrete gold standard methods. This approach was utilised for experiments which apply the Ulm-TSST datasets [24]. The RAAW method essentially targets two core issues within emotional gold-standard creation, i) the alignment of ratings based on the delay which is common for continuous annotation, by up to 6-seconds [124] ii) the disagreement or poor quality of a single rating .

This first issue is tackled by a calculation of Generalised Canonical Time Warping (GCTW), which is an extension of the well-known Dynamic Time Warping (DTW) method. DTW is an alignment approach which implements a distance metric to add elastic properties, computing the best global alignment between signals based on a one-to-many mapping from data points. An extension of DTW is Canonical Time Warping (CTW) [125], which in addition to DTW integrates Canonical Correlation Analysis (CCA) [126], a method for extracting shared features from two multivariate data points. CTW was first applied in the context of computational analysis of human behaviour as a method to align human motion from a multimodal time series [125]. Combining CCA with DTW allows for a more flexible time-warping that handles local spatial deformations of a time series. The CTW was further extended in [127] to GCTW, and this enables multiple sequences (CTW should be done in a pair-wise fashion) to be fused in a computationally efficient way via a reduction of the quadratic to linear complexity. Essentially, RAAW calculates both the alignment, with GCTW and the annotator weighting, with EWE. First, the alignment with GCTW is performed, and then within the RAAW calculation, the similarity (or agreement) is also reported utilising CCC. If in the case of a negative correlation, the signal is excluded from the final step, which is the weighted EWE fusion.

## 3.2 Representations of Audio

In a machine learning paradigm, there are several ways in which the audio signal can be represented. Although great strides have been made, the problem of representation is an ongoing area of research that remains a challenge in the computer audition community. This is particularly relevant as computational capacity becomes a crucial aspect of embedded intelligent systems.

It is important to note, that end-to-end methods will typically bypass the need for feature-based representations and learn directly from the raw audio signal, however depending on the task itself, it may be fruitful to explore lower-level representations which are tailored to the domain. To obtain these lower-level feature-based representations of audio, a transformation of the audio-signal into the frequency domain must first be performed commonly via the calculation of an Fast Fourier Transform (FFT) over the given sequence (for further detail on the process for applying the mathematical function of an FFT see [128]. As audio is a time-dependent modality, and an FFT will not give information on how frequency is changing over time, only its magnitude, and so the Short-time Fourier transform (STFT) is then applied to obtain frequency magnitude information over time, where a frame (or window), and degree of overlap is set statically, moving forward through the signal at a hop size (step to next FFT).

Lower-level representations can then be extracted from this frequency over time representation either over the entire sequence or again by applying a sliding-window at a static hop size. As with all other aspects of the machine learning pipeline, the optimisation of frame and hop size at this point is a critical aspect to consider. It is also worth noting that for speech-driven tasks, the frame-size can be based on the length of an utterance or with consideration to speech pauses or voiced activations of speech (by applying a Voice Activity Detection (VAD) algorithm). This type of segmentation may be particularly relevant and more meaningful to the phenomena in question, as this process reduces the impact of other acoustic activity in the environment.

### 3.2.1 Acoustic Low-Level Descriptors

Acoustic Low-Level Descriptors (LLDs) are, as the name suggests, low-level (i. e., close to the signal) representations of the audio signal extracted at discrete intervals over time. There are several LLDs which are commonly used in the context of speech, particularly computational paralinguistics analysis and recognition, namely prosodic-based features e. g., speech rate, and acoustic-based, e. g., cepstral or spectral derived features [1].

In Chapter 4, a number of the experiments utilise feature sets based on extracted LLDs, and it is also common for LLDs to be used directly with great success [9]. Prosodic features are the core of speech-based tasks and include features that are aimed at modelling the suprasegmental aspects of speech, and from a linguistic and phonetics perspective, this includes; pitch ($F_0$), loudness and rhythm. However, several acoustic LLDs, not rooted in the etymology of linguistics or phonetics, have been found also to model suprasegmental properties of speech and most computational paralinguistics literature will also include these within the prosodic feature grouping [129]. Here a description of four of the more common LLDs is given, particularly those which are referenced in the acoustic analysis conducted in Chapter 4 Section 4.2:

*Pitch ($F_0$):* $F_0$ measured in Hertz (Hz), describes the pitch contour of the fundamental frequency (the lowest frequency in the harmonics of speech). Extracting $F_0$ from the speech signal can be performed utilising a pitch detection algorithm, and several well established methods can be applied for this [130]. Due to its lower computational capacity, the Average Magnitude Difference Function (AMDF) [131] is a standard method for many applications, with several extensions of AMDF being published which aim to improve on what is referred to as the 'falling tendency' [132]. When utilising the $F_0$ contours, it is common to calculate segmental based difference in distribution of the signal,s e. g., the mean, median or skewness. Concerning pitch are the formants of speech, which represent frequency peaks within the frequency spectrum, and it is considered that each formant refers to a resonant activation in the vocal tract, particularly the tongue (F1-F2) and lips (F3) [133]. Pitch is known to alter when a number of states of wellbeing being are influx, particularly conditions such as depression which is known to have a lower $F_0$ during periods of depressions [134].

*Intensity / Loudness:* As with pitch, the intensity contour can be extracted directly from the audio signal. Loudness, i. e., sound intensity, is the degree of sound pressure intensity which is being sensed, and the relative logarithmic measure for this is known as Sound Pressure Level (SPL) [135], however, this should be considered as a looser proxy for speech intensity, as the intensity would be proportionally impacted by all other acoustic events. As with pitch, lower intensity is also known to be linked to poorer states of wellbeing including general sadness and depression [136].

*Duration:* The rhythm or duration of a speech utterance can be automatically calculated in based on a summation of intensity and refers to the activation of speech, including voiced, unvoiced and silent segments [137]. From this descriptor, aspects such as

speech rate can also be observed. Speech rate is beneficial in the context of emotional wellbeing states, as it is known to alter, particularly during emotional expression [138].

*Harmonic-to-Noise Ratio (HNR):* The HNR is metric for the ratio between periodic and non-periodic components with a segment of voiced speech [139]. The overall HNR value for a given signal will vary as different vocal tract activations effect the amplitude of each of the harmonics. There are two components, the vibration of the vocal cords and the glottal noise expressed in decibels (dB), it is the relationship between these components which denotes the speech quality. HNR is discussed in the literature as an indication of dysphonia i. e., disorder of the voice [140].

Considered as part of general acoustic LLDs [80], and probably the most well-establish method for representing the audio signal are the Mel-Frequency Cepstral Coefficients (MFCCs). Proposed initially in 1976 [141], MFCCs are still applied for state-of-the-art modelling today and remain a robust baseline for many computer audition tasks. The MFCCs are highly motivated by human-hearing and are based on the extraction of *cepstral coefficients* from a given number of filter banks, spaced based on the mel-scale [142]. As with most other approaches, the speech signal is first transformed to the frequency domain, and filter banks of a set amount of frequency bands are applied. These filter banks are based on the logarithmically spaced mel-scale, with an intensity weighting of the frequencies to represent the perceptual nature of audio better. As with the purely prosodic features typically, the zero-coefficient (log power) and segmental based difference e. g., delta, and delta-delta, will be calculated for MFCCs.

#### 3.2.1.1 Hand-Crafted Features Sets

Based on a number of the aforementioned LLDs, hand-crafted feature sets have been compiled to 'brute-force' the representation of the target phenomena, with particular success for computational paralinguistics targets [76, 49]. Such sets can be generated in a pseudo-manual way with python-based toolkits such as Librosa [143], or several established features sets can be extracted with the well-known openSMILE toolkit [144]. The ComParE set is a widely used hand-crafted set that can be extracted with openSMILE, and remains competitive for SER based tasks [145]. The latest ComParE set from 2016 [76] consists of 6 373 functionals which have been calculated based on a set 65 LLDs from 4 core groups of acoustic LLDs, prosodic, spectral, cepstral and voice quality. An overview of the LLDs for this set is given in Table 3.4, with further details for spectral and voice quality LLDs, or other which have not been described herein given in [142].

Table 3.3: The 65 LLDs for the 6 373 derived functionals of the ComParE set.

| **4 Energy-based LLDs** | **Group** |
|---|---|
| Sum of spectrum (loudness) | Prosodic |
| Sum of RASTA-filtered spectrum | Prosodic |
| RMS Energy, Zero-Crossing Rate | Prosodic |
| **55 Spectral LLDs** | **Group** |
| RASTA-filtered spectrum bands. 1–26 | Spectral |
| MFCCs 1–14 | Cepstral |
| Spectral energy 250–650 Hz, 1-4 kHz | Spectral |
| Spectral Roll-Off {0.25, 0.5, 0.75, 0.9} | Spectral |
| Spectral Flux, Centroid, Entropy, Slope | Spectral |
| Psychoacoustic Sharpness, Harmonicity | Spectral |
| Spectral Variance, Skewness, Kurtosis | Spectral |
| **6 Voicing-related LLDs** | **Group** |
| $F_0$ (SHS & Viterbi smoothing) | Prosodic |
| Probability of voicing | Voice Quality |
| log. HNR, Jitter (local & $\delta$), Shimmer (local) | Voice Quality |

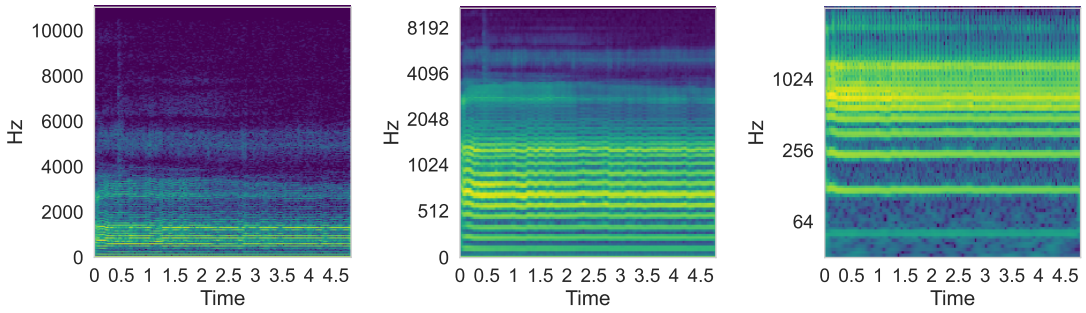
Table 3.4: The 25 LLDs for the 88 derived functionals of the eGeMAPS set.

| **3 Energy / Amplitude-related LLDs** | **Group** |
|---|---|
| Sum of spectrum (loudness) | Prosodic |
| logarithmic HNR, shimmer (local) | Voice Quality |
| **15 Spectral LLDs** | **Group** |
| $\alpha$ ratio, 50-1000 Hz, 1-5 kHz | Spectral |
| Hammarberg index [146] | Spectral |
| Spectral slop, 0-500 Hz, 0-1 kHz, Flux | Spectral |
| Formant 1, 2, 3 relative energy | Voice Quality |
| MFCCs 1-4 | Spectral |
| Harmonic difference H1-H2, H1-A3 | Cepstral |
| **8 Frequency-related LLDs** | **Group** |
| $F_0$ (semi-tone) | Prosodic |
| Formants 1,2,3 (frequency) | Voice Quality |
| Formants 1,2,3 (bandwidth) | Voice Quality |
| Jitter (local) | Voice Quality |

As well as the ComParE set and a number of others, there is also the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) set [51], which was specifically designed with emotional speech in mind. The LLDs that are selected for the eGeMAPS set were chosen by the authors through a discussion with interdisciplinary speech scientists and is justified by three core criteria i) the ability of a particular LLDs to model a physiological change in vocal production during an emotional state ii) the success that the particular parameter had in the context of machine learning from speech iii) and the theoretical significance of that feature as it pertains to alternative research in the acoustics of emotional speech . The eGeMAPS set is much smaller than ComParE and consists of 88 functionals, which are derived from 25 LLDs (see Table 3.4, for an overview).

### 3.2.2 Image-Based Learnt Representations

Although remaining robust and competitive for many computational paralinguistics-based machine learning tasks, the hand-crafted sets are, in modern machine learning, often criticised due to their heavy need for human intervention [40]. An alternative method for extracting audio representations is via image-based pre-trained networks and audio plots. Commonly spectrograms are used, but depending on the task there are a number of alternative audio plots which can be extracted (see Figure 3.5). As with the hand-crafted feature sets, several image derived feature sets are extracted and compared in the experiments of Chapter 4, particularly in relation to anxiety in speech, via a comparison of effect size between grouped speech-classes in Section 4.2.




(a) Spectrogram (b) Mel-Spectrogram (c) Constant-Q Transform

Figure 3.5: A selection of audio plots from a sample vocalisation of a sustained vowel [‘a’]. Of note, it can be seen that selecting the most appropriate audio plot can be extremely meaningful for highlighting certain aspects of a speech signal, e. g., the harmonics of speech.

To learn meaningful representation from audio plots, several unsupervised strategies which utilise deep learning have been developed which essentially mimic well-established methods in the vision domain. For audio specifically these include, VGGish [53] DeepSpectrum [85] and auDeep [147]. At a first step, these approaches require an audio plot and this is highly dependent on the given task, how in Figure 3.5, those which may applicable for speech-based tasks are shown.

The DeepSpectrum toolkit is one method which has been utilised for several computational paralinguistics tasks [28, 19] and applies an image-based deep learning approach (see Figure 3.6 for an overview of this method). This method essentially has three core-processes i) the extraction of two-dimensional representations of the audio signal, namely the most meaningful audio plot for the tasks ii) audio plots are then sent to a pre-trained Convolutional Neural Network (CNN), with options including, AlexNet [148] iii) after the plots have been fed through the CNN, activations from the fully-connected layers they are then extracted as feature vectors . Given the inherent visual dominance of this method, it has been noted by the toolkit authors themselves [149] that the convolutional layers of the CNN can provide a strong indication of the locality of pixel dependencies, and so in the context of the speech, the more prominent a specific phenomena is visually within the audio plot, the more applicable the DeepSpectrum representation would be.

Similarly to DeepSpectrum, is the VGGish model for audio feature extraction [53]. VGGish is an audio adaption of the vision focused VGG16 architecture which is applied to image-based tasks [150], and was first introduced as a baseline for the aforementioned AudioSet. The framework provided for extracting VGGish features, first extracts spectrogram-based representations which are then fed to a CNN which was pre-trained on clips from the large-scale YouTube 8M dataset [84]. The core nature of these features is more focused on

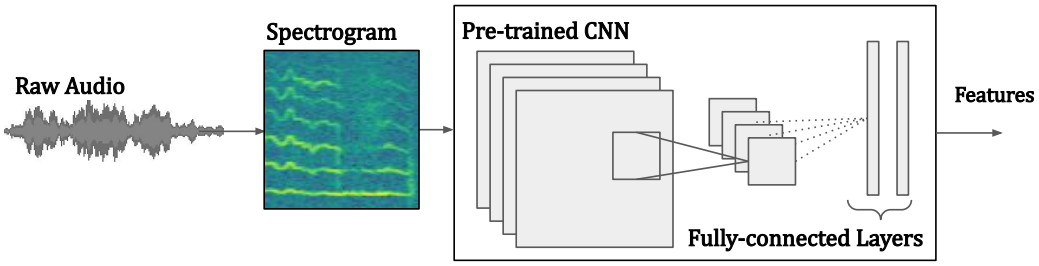



Figure 3.6: An illustrated (adapted from [149]) overview of the DeepSpectrum method for a pre-trained CNN-bassed feature extraction from audio plots. Where the raw audio is first transformed to an audio plot, which is then fed to a pre-trained CNN and features are extracted from the activations of the networks fully-connected layers.

audio event detection, however, there have been a number of successful application for these features in the domain of SER [15].

### 3.2.3 Fusion Strategies

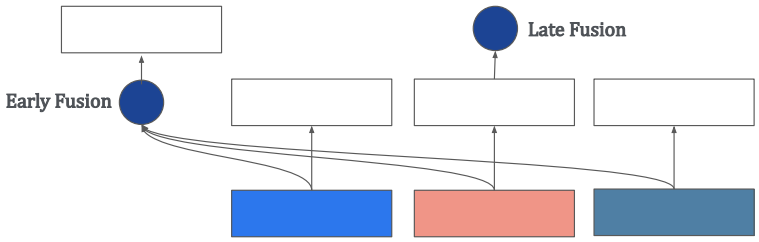



Figure 3.7: An illustration of the two-types of the fusion strategy applied in this thesis. Figure is an adaption of the illustration given in [151].

In many cases for computational paralinguistics tasks it is beneficial to explore the fusion of several feature sets [49], either other acoustics sets, or for exploring the benefit of audio plus other modalities (i. e., multimodal), as well as multiple modelling approaches. There have been a number of works which have found that fusing audio with video-derived features can be very beneficial [15], and this is further explored in the experiments of Chapter 4 Section 4.1. This is particularly the case for in-the-wild scenarios, as the individual modalities may be at times obscured, and fusing feature representations allows for an improved overall representation of the given phenomena. Applied within this in this thesis, are two typical approaches for fusion (see Figure 3.7):

*Early fusion:* Also known as feature-level fusion, early fusion refers to the concatenation of the individuals features sets into a single large feature set prior to model training [25].

In this scenario it is important that the individual feature sets relate to each other, i. e., the sampling rate, and to adjust this interpolation, or downsampling can be applied, to increase or decrease the data frequency [30].

*Late fusion:* Also referred to as decision-level fusion, late fusion is the process of combining predictions from the output of multiple models, one for each feature type. The approach for this is based on a defined rule, which can differ depending on the output as either a regression or classification task. Where majority vote is common for classification, and a weighted average of the best performing predictions can be taken for regression [19]. However, for regression-based tasks, late fusion can also include an additional linear regression model trained on the predictions of each model to learn the weights which would then be combined by the weighted average [15, 24].

## 3.3 Machine Learning

Within machine learning applying the most appropriate algorithm to model a given phenomena – based on a training set of examples – is a critical decision, and as described earlier in Chapter 2 this will mostly depend on the nature of the task as either a classification or regression problem. However, within this there are either supervised or unsupervised algorithms [152].

- A *supervised* algorithm is as the name may suggest, an algorithm which has been provided information on the values of their training input and are essentially trying to finding a mapping between this input $x$ and a given set of labels $y$, typically by estimating the probability $p(y|x)$ [152].
- An *unsupervised* algorithm has no prior information about its input $x$ and will attempt to learn patterns within the embeddings space. In this case the algorithm is attempting to learn the probability distribution $p(x)$. Clustering algorithms are a form of unsupervised learning, which attempt to build clusters of similar examples.

The experiments described in Chapter 4 are primarily applying supervised algorithms and in that sense tackling either classification or regression based problems. Herein, the fundamentals for all technical strategies applied to the emotional wellbeing targets is given.

### 3.3.1 Recognition

In the following, a series of algorithms are introduced which have been applied within the context of the following experiments and proposed as suitable strategies for audio-based

recognition of states of emotional wellbeing. The algorithms discussed can be applied to learn patterns within the audio data in either a supervised or unsupervised way, and so the use of the term recognition in this context relates to learning patterns within a data input.

#### 3.3.1.1 Support Vector Machines

The Support Vector Machine (SVM) is a supervised machine learning algorithm [153] which has been found to be a versatile and robust baseline algorithm for an array of computational paralinguistic tasks [1], and remains a standard for a number of challenge baselines where reproducibility is vital or the dataset size is not extensive (see Chapter 2). This algorithm is proposed for use in a number of the experiments in Chapter 4, and performance for the SVM is compared to deeper machine learning algorithms in Chapter 4 Section 4.1.

An SVM is a decision machine, or maximum-margin classifier and does not provide probabilities, but rather a class-decision on a given sample. The SVM essentially functions by calculating what is known as a *hyperplane* between the classes of a given training set. To do this the SVM attempts to find the maximum margin between the closest vectors of a given class (known as the support vectors), then calculates decision boundaries based on this. In its native state the SVM can only be applied to binary-class problems, however to apply an SVM to a multi-class problem, there are two methods known as one-versus-all, in which each class is classified against all other classes, or alternatively, the one-versus-one strategy in which pairs are classified against one another. The SVM can also be adapted to function as a regressor, where the target is a real numbers, and is known as Support Vector Regression (SVR) in that case. Unlike traditional linear regression, an SVR allows for more margin of error, as it introduces an $\varepsilon$-insensitive region, which is calculated by the application of a symmetrical loss function (commonly, $\varepsilon$-insensitive loss [154]) which equally penalises values which are to low or too high outside of what is described as an $\varepsilon$*-tube* around the estimated function [155]. In both cases for either regression or classification, the value of cost $C$ should be optimised, which essentially refers to the tolerance for values to be excluded or included within the decision boundaries $\varepsilon$.

#### 3.3.1.2 Artificial Neural Networks

A core part of state-of-the-art machine learning is the Artificial Neural Network (ANN). An ANN is commonly described as being inspired by the workings of the human-brain and is essentially based upon a complex set of neurons which receive a set of inputs, and outputs a single value [156]. The Feed-Forward Neural Network (F-FNN) is a starting point for understanding this concept (see Figure 3.8). The F-FNN is a type of ANN and refers to

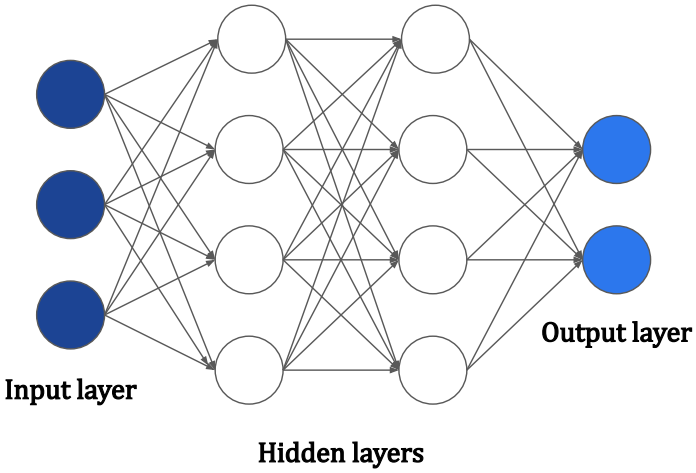



Figure 3.8: A simple illustration of a F-FNN with an input layer, which feeds into two fully-connected hidden layers, and an output layer which provides the target prediction.

a network in which the input data moves forward in one way, through what are known as hidden layers. Where this becomes a Deep Neural Network (DNN) or deep learning, in general, is when there is more than one hidden layer, in other words, more depth to the structure. Essentially, any ANN is performing a step-based optimisation and consists of several layers of nodes (i. e., perceptrons), with summation of weights and a bias passed to the input of an activation function which manages the final output. In other words, a given input node has a weight $w$ which is optimised during the training process, representing the importance of that node, and the bias $b$ is a constant value applied to shift the output before the activation function. Once the input reaches the last layer, the final prediction is made, and a loss function will calculate the error between the predicted values $\hat{y}$ and the true value $y$.

A *loss function* essentially calculated the error between $\hat{y}$ and $y$, and minimising the loss value is a core aspect of the ANN. Many loss functions can be applied during training and are reduced to obtain what is known as the local minimum of the function. Common loss functions include, Cross-Entropy (CE) in the case of classification, and Mean Absolute Error (MAE) for regression, although this can be adjusted depending on the task itself, with Concordance Correlation Coefficient (CCC) commonly applied for emotion-based targets [60]. The loss is minimised via the optimisation method, gradient descent. For this, the differentiation algorithm backward propagation of the errors (back-propagation) is performed to calculate the gradient of the loss function, applying the chain rule to adjust weights and bias for the model by iteratively going forward and then back through the network. To reduce computational expense of this, the processes of gradient descent is performed iteratively on stochastic batches of a given number of samples from the training set, therefore known as stochastic gradient descent. One option is the extension of stochastic gradient descent known

as the Adam optimiser [157], which is commonly applied in deep learning, and is essentially combining the advantages of other stochastic gradient descent extensions, e. g., the Adaptive Gradient Algorithm (AdaGrad) or Root Mean Square Propagation (RMSProp).

The *activation function* is a crucial aspect of any ANN and refers to the function which can be applied to obtain the precise output of a given node (or perceptron). There are a number of both linear and non-linear activation functions which should be selected based on the hidden and output layers. Depending on the type of target, commonly applied activations functions include:

*Sigmoid*, or the logistic sigmoid function, would typically be used only for the very last layer to predict the final value of $y$. The function is limited to the range of [0, 1]:

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}. \tag{3.2}$$

*Hyperbolic tangent function* (tanh) is an extension of sigmoid, and allows for negative values, scaling and shifting the sigmoid function within the range of [-1, 1]:

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1. \tag{3.3}$$

*Rectified linear unit* (ReLU) is a very popularly applied activation function for a number of reasons including it efficient computing ability:

$$\text{ReLU}(x) = \max(0,x). \tag{3.4}$$

The output of ReLU is positively unbounded, which means the output can grow exponentially, and an additional regularisation of the weights may be beneficial.

*Softmax* is a well-known activation function which was developed for multiclass classification. All neurons $K$ of the final layer are transformed into probabilities between [0, 1] for each class:

$$\text{Softmax}(x) = \frac{\mathrm{e}^{x}}{\sum_{k=1}^{K} \mathrm{e}^{x_k}}. \tag{3.5}$$

Activation functions can be applied to different layers within a network to optimise performance, and as well as this there are a number of methods which can be utilised to optimise a network known as hyperparameters. These include but are not limited to:

- The number of *epochs* refers to the number of iterations for which the data should be passed through the network. In other words a forward and backward pass through all of the data.

- The *learning rate* controls the number of updates to the weights at each iteration, i. e., the rate to which a DNN adapts to the task and reaches global minimum. A lower learning rate will require more epochs to reach the global minimum, as more slight adaptations are made.

- The *batch size* refers to the number of samples from the training set which will be propagated through the network in a given iteration. The size of the batch will effect the training time, and the overall memory allocation required.

Given the exponentially large size of a DNN the network can become uncontrollable, and what is known as overfitting i. e., where a network learns the data in a completely unrealistic way, can occur quickly. Given the small amount of data which is typical for computational paralinguistic tasks, the phenomena of overfitting is even more likely, so as well as having more data, there are many specific methods in deep learning, known as regularisation strategies, which attempt to produce more robust results including:

- *Dropout* is a standard method applied to a ANN [158], and performs model averaging, aiming to prevent complex co-adaptations in a network. Essentially it does this by thinning the network and dropping neurons from within a hidden layer.

- *Early stopping*, is both a regularisation and efficiency step, which can be put in place to stop the training process once the model is no longer improving (i. e., able to minimise the loss) [159]. Typically, what is known as *patience* would be set in this case to ensure that the model is not stopped too early as sometimes may recover.

#### 3.3.1.3 Recurrent Neural Networks

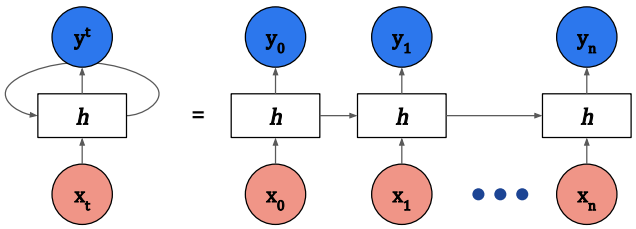



Figure 3.9: A depiction of an Recurrent Neural Network (RNN) (left side), where the input $x$ passes to the RNN cell $h$ and the output is given as $y$. The right side of the figure, shows an example of the RNN unrolled, in relation to BPTT process.

Unlike the F-FNN which is primarily aimed at a static inputs and spatial tasks, the RNN has an internal memory, which allows for the output to be used again as the input, and is,

therefore, able to capture temporal dependencies. In other words, where the F-FNN can only map a *one-to-one* sequence, the RNN is able to map *one-to-many*, *many-to-one*, and *many-to-many*. Consequently, it is instrumental in the context of time-dependent audio tasks, and has been successfully applied to an array of speech-based tasks including, language understanding [160], and speech emotion recognition [161]. In the context of these this thesis the RNN is the basis for many of the experiments described in Chapter 4, and as mentioned earlier and LSTM-based RNN is explicitly compared to the performance of the traditional SVR for its ability to recognise markers of stress in the experiments of Section 4.1.

The RNN, consists of a set of recurrent layers, which contain recurrent perceptrons. The core difference between a feed-forward perceptron and a recurrent perceptron is that the output is again fed as an input, this is known as a *feed-back* connection which, when combined with several other recurrent perceptrons, makes up an RNN. An RNN is able to handle sequential data of a given length, in what is referred to as *segments*, this is particularly suitable for speech data, as words of a sentence within a particular context can be processed together, allowing for a deeper more context-driven learning by the network. Given this sequential nature, for a RNN the process of back-propagation is extended to Back-propagation Through Time (BPTT) [162], which is a slight alteration, and essentially means *unfolding* the RNN and for each input step there is one copy of the network and one output, the gradient is then calculated at each step and summed, (see Figure 3.9 for an illustration of this concept). The network is then folded up again, and weights are updated. This is the process that can make the RNN computationally expensive, particularly when larger segments are being processed, and for this a truncated BPTT can be applied, which only performs the BPTT periodically and on a selection of input steps and not all.

The problem with the RNN on its own is that as weights are updated using BPTT, the gradient can decrease exponentially as back-propagation continues, meaning that the weight is then not updated, and the effect of earlier inputs is not learnt. This is known as the *vanishing gradient problem*, and to address this, two strategies can be applied, e. g., Long Short-Term Memory (LSTM) or Gated Recurrent unit (GRU) cells. The LSTM, is seemly more popular for computer audition tasks, and is proposed for all of the experiments in which an RNN-based network is utilised in Chapter 4 – for this reason, only the LSTM will be described herein, although the GRU is only a slight adaption and in many cases the two can be used interchangeably with minimal effect on results.

The LSTM cell consists of a series of additional gates (i. e., independent units with their own weights and bias); these include the *forget*, *input*, and *output* gates (see Figure 3.10). Sequential data is passed through each of the gates, and the state of the previous time-step and current input are used as input to each gate, which contain an activation function $\sigma$.

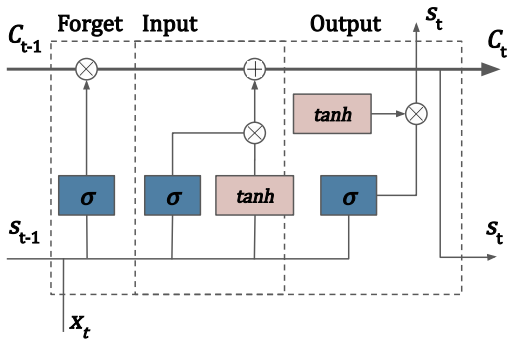



Figure 3.10: An overview of an LSTM cell. Where the input $x$ is passing through each the three gates, forget, input and output. The state of the cell is indicated as $C_t$, in which a point-wise multiplication or addition with the output of each gate $\sigma$ is calculated. The previous state is as $s_t - 1$, which is concatenated to $x$, and the output to the next layer is indicated as $s_t$.

Essentially the forget gate controls which information from the previous state should be forgotten or kept from the initial units state, the input gate decided if the last cell state is to be stored, and output gate controls how much of the input should be outputted to the new state.

#### 3.3.1.4 Attention Mechanisms

Attention mechanisms can be integrated with an architecture such as an Long Short-Term Memory Recurrent Neural Network (LSTM-RNN), and are named based on the idea of e. g., cognitive attention. The attention mechanism was initially introduced to improve sequence-to-sequence modelling in the area of Natural Language Processing (NLP) [163], and are therefore inherently applicable to time-dependent tasks. There have since been several alternative attention methods proposed, which can be utilised in either a shallow or deep way in a deep learning architecture e. g., an LSTM-based network [164]. In general, attention mechanisms allow for a broader consideration of past and future data points within a sequence. Further, given that attention is calculated based on the weighting of the input against all others in the sequence, the activations from an attention layer can be extracted to provide an understanding of the more meaningful aspects learnt by the network [165].

As mentioned, attention mechanisms are prevalent for sequential based tasks as they allow for a consideration of the larger context. In the experiments of Chapter 4 Section 4.3.1, self-attention is integrated with the LSTM-RNN architecture for a SER task, as it has been found to be applicable for this [66]. Occasionally self-attention is referred to as intra-attention, and this is one type of attention that is the building block for the recently popular Transformers network [17]. Further details of the self-attention mechanism are given in [17],

but fundamentally, the self-attention functions by calculating a dot product of a given vector $x_i$ against its neighbouring vectors $x_1 ... x_n$, with each $x_i$ given three times each with a differing weight, either as query (the previous compressed output), or key, value pairs, combined via a softmax function to ensure positive weights for each vector.

#### 3.3.1.5 Convolutional Neural Networks

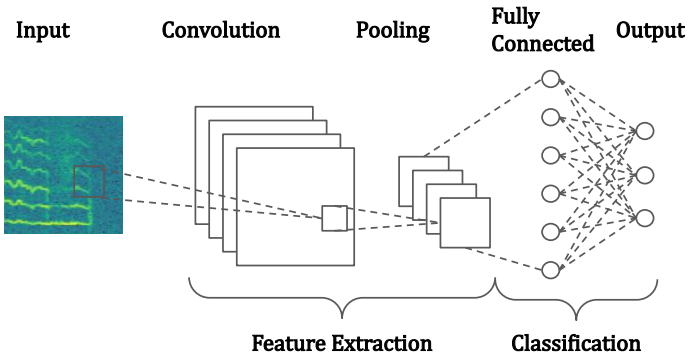



Figure 3.11: An overview of CNN architecture. Where a spectrogram image is inputted to the convolutional layers, which perform an abstraction of the input, and then the pooling layers downsample the representations further, and finally those activations are passed to a classification algorithm in this case F-FNN.

The Convolutional Neural Network (CNN) is a type of ANN which has become one of the most common deep learning networks for classification tasks. Typically a CNN takes a two-dimensional image as input and has found great prominence, where reduction of the complex spatial relationships is needed. In recent years the CNN has been successfully applied in general for two-dimensional audio tasks, where the input is an audio plot, which represents the time and frequency domain of audio. In this way, the CNN is applied as part of the feature extraction process described for the aforementioned image-based methods in Section 3.2.2, and is also applied as part of the architecture proposed for interpreting generated audio in Chapter 4 Section 4.4. The CNN consists of a set of convolutional layers and pooling layers to extract features from the input image, and a fully-connected output layer which will classify the extracted activations (see Figure 3.11). Although there are additionally one or three-dimensional CNN methods [166], here the two-dimensional approach is described, as would be applicable for audio-plots.

A convolutional layer consists of a kernel or filters that slide over the input image at a given dimension, calculates an element-wise multiplication at each step, and reduces that kernel to a single output, resulting in an abstracted representation of the original input image. Proceeding this, an activation function (e. g., ReLu) applies element-wise non-linearity, and

then a pooling layer (either average or max pooling), is placed after the convolutional layers. This layer is applied to reduce the dimensionality of the representations progressively, and additional pooling layers can be used as a regularisation method [166]. The output from the pooling layer is then passed to a fully connected F-FNN to perform the final classification.

#### 3.3.1.6 Prototypical Networks

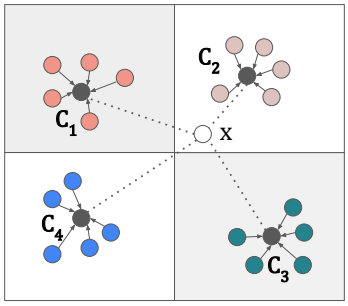



Figure 3.12: An illustrative example (adapted from [167]) of a prototypical embedding space in a few-shot learning scenario. Where a class prototype $c_k$ is computed from the mean of the examples within the embedding space of that class, and the distance of the input $x$ is measured to assign a class prediction.

The prototypical network is a network which was introduced by [167], and essentially works on the assumption that there exists a prototypical representation of a given class (see Figure 3.12). The network is primarily applied to few-shot and zero-shot learning classification of images [167], text [168], and audio [169]. This network is proposed as a strategy for assisting in the interpretation of generation audio as it pertains to the source training data, for which the experiments are details in Chapter 4 Section 4.4.

The terminology utilised for the prototypical network differs slightly, and as mentioned, a prototypical network is searching for a prototypical (or stereotypical) class $k$ within an embedding space, from data points provided as a *support* set $S_k$ – an input $\mathbf{x}_i$ and labels $\mathbf{y}_i$ for each class $k$, function as an anchor for the class-prototypes $\mathbf{c}_k$ – and the distance of these is compared to a *query* set – as with $S_k$, excluding $\mathbf{y}_i$. Essentially, a prototypical network learns an embedding function $f_\phi$, which maps the input to a $N$-dimensional embedding space. The prototype for $k$ is calculated from the average of the support set embedding as:

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i). \tag{3.6}$$

Further, the euclidean distance $d$ between a class-prototype and the embedding of a *query* is input to a softmax function, allowing the network to additionally serve as a classifier. For

the embedding function ($f_\phi$), which learns data representations based on the classes, in [167] the authors initially apply CNN-based architecture. There were also a set of *design choices* introduce by [167], including the episodic training style, which is essentially mini-batch of a sub-sampled number of classes and data points, and insures balance across the classes.

#### 3.3.1.7 Evaluation Metrics

Several metrics can be applied to evaluate the performance of a recognition model, by comparing the final predicted labels in a discrete or continuous state to an actual label for the same input taken from the testing partitions. Usually, it depends on the nature of the task, either classification or regression, subjective or objective, as to why a suitable evaluation metric would be chosen.

For a classification paradigm, one way to interpret class-based tasks is through visualisation of the individual class recall via a *confusion matrix*. A confusion matrix typically depicts a binary task, however, it can be extended to observe multi-class performance. The confusion matrix, reports accuracy of a given class it pertains to the rate of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) for that given class. In the case of a multi-class problem, the FP and FN would be extended across the other classes. These concepts are essential when calculating a given evaluation metric in the context of classification, where the prediction is either correct or not.

*Accuracy* is a fundamental and common metric which can be calculated as:

$$\text{Accuracy} = \frac{\sum \textit{correct}}{\sum \textit{observations}} = \frac{TP+TN}{TP+TN+FN+FN}. \tag{3.7}$$

Accuracy is brute-force, and as discussed earlier in Chapter 2 class imbalance in common in this domain and so reporting accuracy alone can misrepresent true performance.

*Precision* and *recall* can be applied to consider performance of a single class as:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{3.8}$$

$$\text{Recall} = \frac{TP}{TP+FN}. \tag{3.9}$$

*F-score (*$\text{F}_1$*)*, is the harmonic mean of precision and recall:

$$\text{F}_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{3.10}$$

*Unweighted Average Recall (UAR)* is applied to computational paralinguistic tasks as this does not consider the frequency of each class and is therefore handling the problem of class imbalance. Essentially UAR is the sum of individual recalls for each class $n$ and a calculation of the average for the total number of classes, denoted as:

$$\text{UAR} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_n}{TP_n + FN_n}. \tag{3.11}$$

For regression tasks, there are many ways to evaluate the relationship between the actual value $y$ and the predicted value $\hat{y}$, commonly via correlation, or an error-based metric. The choice of either a correlation or an error-based metric depends on the task itself, however the latter is typically applied to objective tasks e. g., heartbeats per minute, and correlation is more suitable for subjective tasks, e. g., emotion recognition.

*Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE)* are all error rates that can be commonly applied to evaluate a regression model. Often reporting all will allow for more insight, however RMSE is common for linear regression based tasks as its gradient is linear.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|, \tag{3.12}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2, \tag{3.13}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}. \tag{3.14}$$

*The Pearson correlation coefficient (r)* is a metric used to observe the linear correlation between two continuous arrays. Pearson correlation ($r$) is expressed as:

$$\text{r} = \frac{\sum(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum(x_i - \hat{x})^2 \sum(y_i - \hat{y})^2}}, \tag{3.15}$$

where in this case $x$ refers to the actual label value and $y$ the prediction and $\hat{x}$, and $\hat{y}$ are the mean of the values for that array.

*The Spearman's rank correlation coefficient ($\rho$)* , is essentially the ranked version of the Pearson coefficient, and is more common for values derived from an ordinal value.

First, the two arrays being compared are ranked as $r_x$ and $r_y$, and the standard deviations $\sigma$ of the ranked data is then divided by its covariance:

$$\rho r_x, r_y = \frac{cov(r_x, r_y)}{\sigma\ r_x \sigma\ r_y}. \tag{3.16}$$

*The Concordance Correlation Coefficient (CCC)* is another evaluation metric which is commonly used in emotion recognition due to the implicit scaling variance between these subjective rating. CCC can be which can be expressed as:

$$\text{CCC} = \frac{2\sigma_{12}}{(\mu_1 - \mu_2)^2 + \sigma^2{}_1 + \sigma^2{}_2}, \tag{3.17}$$

it is assumed here that the two variables follow a normal distribution with the mean $\mu_1$ $\mu_2$ and standard deviation $\sigma_1$ $\sigma_2$ over the covariance of the two signals as $2\sigma_{12}$.

## 3.3.2 Generation

Generating artificial audio based on a set of conditions defined by the training set has become a more needed aspect for computer audition, particularly computational paralinguistic focused tasks, as data is sparse and expensive to gather. However, a number of the methods described herein are also applicable to the field of Text-to-Speech (TTS), although in this case, speech intelligibility is naturally the first priority. Unlike TTS, when utilising generation methods to tackle data scarcity via data augmentation of existing data, especially in regards to paralinguistics, intelligibility is less critical, as the machine is the only one listening, and the phenomena is less related to lexical meaning. There are a number of machine learning strategies which can be applied to generate audio data. Most prominently those known as deep generative models, which essentially provide a representation of a probability distributions over multiple variables [152]. Such networks include Variational Autoencoders (VAE) Generative Adversarial Network (GAN), and Deep Auto-regressive Networks (DARN), and in this section a description of only GANs and DARNs is given, as these have been applied in works by the author, namely [170] and [26] with further detail given in Chapter 4 Section 4.4. For an overview of core aspects of deep learning-based generative models (see [152, 171]).

### 3.3.2.1 Generative Adversarial Networks

The GAN was first introduced in [156], and has been extensively applied to a vast number of tasks across modalities, with some great success in recent years when adapted to the task of generating raw audio [172]. Essentially, a GAN is based on a differential generation

a network, and the core idea of a GAN is to pit two networks against each other in an endeavour to repetitively improve the results (see Figure 3.13). The first network is known as the *generator* and learns to transform any vector that follows a given distribution function, e. g., a uniform distribution, to an output sample that follows the distribution of a given training domain. The second network (the adversary of the generator) is known as the *discriminator* and is essentially a classifier which learns to distinguish between real training data and the samples produced by the generator [156].

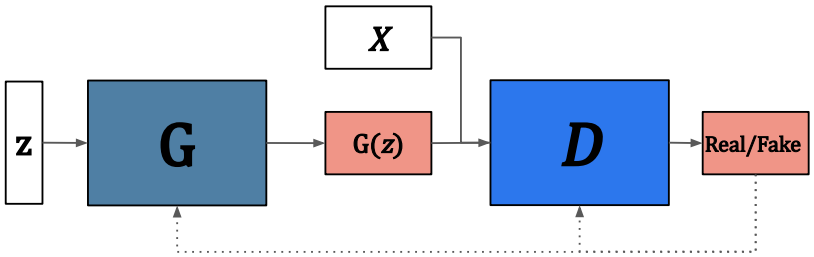



Figure 3.13: An overview of a GAN, where a noise sample $z$ is fed as an input to the Generator $G$, and an real sample $x$ is pit against the generated sample $G(z)$, by the discriminator $D$. The discriminator then classifies the sample as real or fake, and the weights for the generator and discriminator are then updated accordingly.

The WaveGAN [172] architecture is one of the earliest utilises of the GAN for audio generation based on raw audio. The WaveGAN includes an adapted GAN known as the Deep Convolutional Generative Adversarial Networks (DC-GAN), which essentially incorporates convolutional layers to both the generator and the discriminator network to enable even higher complexity modelling [173]. As the DC-GAN is traditionally focused on image generation, for the WaveGAN aspects are modified to enable audio processing. Mainly this pertains to the dimensionality, in that the two-dimensional up- and downsampling filters are replaced by a one-dimensional equivalent. Deeper details of the functionality of WaveGAN are provided by the authors [172].

#### 3.3.2.2 Deep Auto-Regressive Networks

Unlike a GAN, the Deep Auto-regressive Networks (DARN) is able to be used in the context of generation by sequentially modelling the input and output. Much like an RNN, a DARN is learning from past information Figure 3.14, however unlike the RNN the previous input is not given via a hidden state, but directly as a new input based on the output at the previous step [152]. A DARN is typically a set of convolutional layers, which have auto regressive connections i. e., the generated output becomes the input of the next. This flow is able to continue until eventually the input to the network is no longer the original data at all.

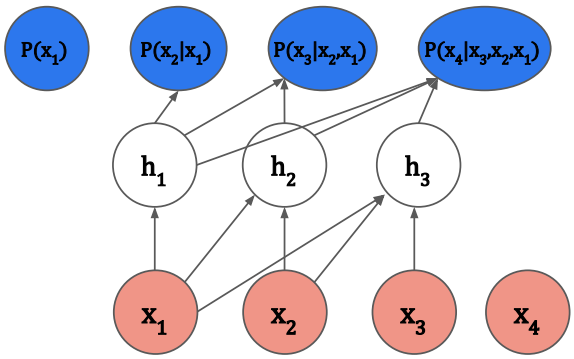



Figure 3.14: An illustration of an auto regressive network, adapted based on the illustration given in [152]. The DARN predicts the i-th input $x_i$ from the previous $x_i - 1$.The parameterisation of the DARN means that groups of hidden units $h_i$ can be reused.

The WaveNet [174] is one of the most well known DARN-based architectures, and was specifically developed for the generation of raw audio by machines. This architecture is essentially an audio implementation of the first of its kind PixelCNN [18]. At its core the WaveNet architecture is generating a given sample of audio $x_t$ and conditioning based on the samples from the previous time steps, and the probability distribution of the waveform $x_1, ..., x_T$ is then a product of a set of conditional probabilities:

$$p(x) = \prod_{t=1}^{T} p_\theta(x_T \ | x_{<t}) \tag{3.18}$$

As a first step, the WaveNet model processes the raw audio, into an 8-bit resolution, with 256 possible values, and the amplitude is then one-hot encoded, and passed through to a set of causal convolutional layers. During the training process, the model predicts audio signal values at each step comparing to the previous step, and applying cross entropy as a loss function. To decrease the computational expense of this, WaveNet applies the method of stacked dilated casual convolutions, which essentially skips nodes of a given hidden layer (see Figure 3.14), this then reduces the receptive field in general. Another method to reduce computational expense is the residual block and parameterised skip connections, which speeds up general convergence, and allows for deeper model training [174].

#### 3.3.2.3 Evaluation of Generated Raw Audio

Evaluating the performance of any generation model is a challenge, but more so in the context of audio. Several qualitative or quantitative approaches have been developed for evaluating image-based generative data, mainly focusing on the GAN [175], and researchers in the audio domain will typically apply these image-based approaches to generate audio data.

Most common and well-known is the inception score [176], which is pre-trained on ImageNet and calculates the logits (e. g., raw unnormalised probabilities) for a given generated sample. The inception score does not make a comparison to the actual source data, and so for this purpose, the Fréchet Inception distance was proposed in [177]. The main limitation for these approaches is, as mentioned in [175], that these two Inception-based approaches are not able to discuss both quality and diversity, e. g., a low inception score or Fréchet distance may be caused by both non-realistic samples, and also samples being too close within the embedding space. Furthermore, although the results from these scores have shown to be similar to human perception, it was noted in [175] that they might be biased towards ImageNet, and therefore less applicable in other domains.

Another, more conventional approach, which is more easily adapted to other domains such as audio, is to use the generated data to augment the original training set and validate the overall performance boost. Given that the generative model is based on the source data training set, the assumption is made that a better performing model will produce data samples within the distribution of the source data. However, this method is a 'black-box', and only limited interpretations can be made about the overall quality of the data [170].

A time-consuming but popular approach and some might argue needed for audio, is to conduct human-perception studies on the generated samples, where the target is compared to the generated sample. However, this is massively time expensive in the context of audio, given its time-based nature. With this in mind, more methods consider many evaluation criteria, e. g., consideration to interpretability and observation of quality and diversity.

Analysis of the embedding space is in general becoming a more deeply investigated area, as this allows for improved data interpretability, which, as mentioned, is a limitation of the inception approaches. One other approach for this is Local Intrinsic Dimensionality (LID) from [178] who introduced the CrossLID method to evaluate the distances of clusters in a 'neighbourhood' within the latent GAN generated space by measuring the distances between two data distributions. Leading from these types of works, in the experiments of Chapter 4 Section 4.4, analysis of the data manifold (i. e., embedding or latent space) is explored utilising the aforementioned Prototypical network Section 3.3.1.6.

# Experiments

Within this chapter a series of experiments are described which focus largely on the evaluation of states of emotional wellbeing, as well as tackling aspects which make modelling this target a challenge. Details of the methodologies for these experiments are described in Chapter 3 in earlier sections, and therefore the theoretical descriptions in this section will be limited. Further to this, where possible it will be highlighted how each experiment relates to each of the research questions described in Chapter 1.

## 4.1 Physiological Markers of Stress from Speech

There is a substantial impact on the body when an individual is stressed, which, if sustained, can result in potentially serious health implications [179]. Physiological signals can provide an objective marker of biological stress e. g., the stress hormone cortisol, however extracting such data can be costly and time-consuming [180]. As stress results in an emotional response by the body and is known to alter expression and an individual's ability to function in a typical manner, speech-based monitoring may capture these changes in various markers and allow for a non-invasive indication of stress. With this in mind, within these set of experiments, the focus is on the recognition of several physiological markers of stress from the audio signal, based mainly on the work conducted by the author and colleagues in [112], and later extended in [31]. Utilising three datasets collected under the Trier Social Stress Test (TSST) scenario, the first experiments focus on the use of audio and multimodal signals to recognise saliva-based sequentially sampled cortisol (the stress hormone). Following this, experiments are conducted to recognised heart rate as BPM, with further analysis on how this may pertain to cortisol groupings. To bring this back to the research questions (RQs) laid out in Chapter 1, these experiments address several RQs in the following manner:

- **RQ-1:** Validating the efficacy of audio-based features for targeting markers of stress, and in turn overall emotional wellbeing.
- **RQ-2:** Exploring the use of audio in a uni vs multimodal setting. Validating the overall benefit of audio for recognising markers of reduced emotional wellbeing.

- **RQ-3:** Exploring the benefit (or not) of combing three datasets which have been collected in differing acoustic environments, under the TSST scenario. Through the combination of these datasets multi-domain analysis possible, and total speakers is increase to more than 100, which is larger than typically available dataset in this area, and may allow for improved generalisation.
- **RQ-4:** To improve the scientific rigour, the data provided was gathered from external partners at several respected psychology research groups. With this in mind these experiments are also a testing bed to explore the benefit that interdisciplinary collaboration can have in regards to computational emotional wellbeing analysis.

Physiological markers are known to relate to the activation of the Hypothalamic Pituitary Adrenal axis (HPA) [181], which is an indication of stress. With this in mind, two core markers are targeted within these experiments 1) sequential saliva-based samples of cortisol 2) continuous Heart Beats per Minute (BPM) . For both experiments the FAU-TSST, Reg-TSST, and Ulm-TSST datasets will be used interchangeably, and for further detail on these, as well as the TSST testing paradigm see Chapter 3. Within these paradigms, several cross-corpus (where possible) and multi-domain experiments are performed for each of the targets and the efficacy of these machine learning approaches for entirely unlabelled data is explored.

### 4.1.1 Data and Procedure

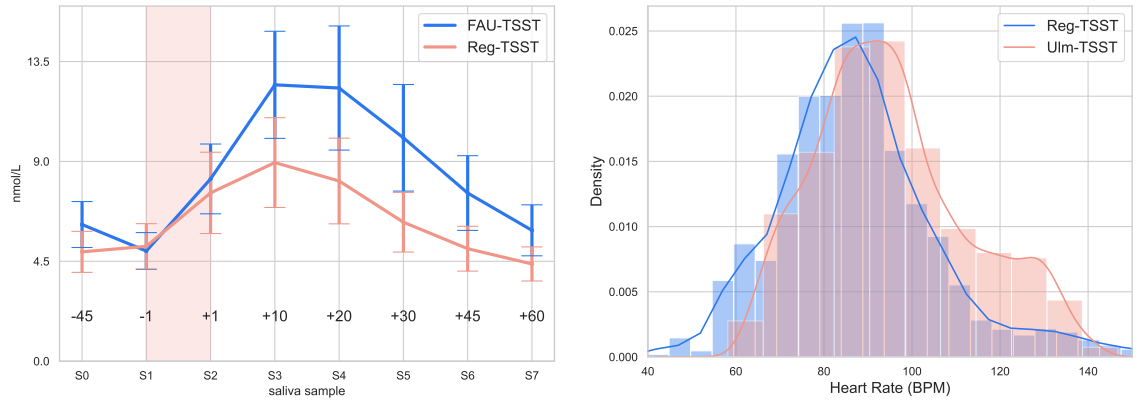



Figure 4.1: The mean of the raw cortisol samples for all speakers of each sequential time step, given in Nanomoles per Litre (nmol /L), for both the FAU-TSST and Reg-TSST. Highlighted is the stress period in grey, with annotations of sample time in minutes (left). The density distribution of the mean BPM across all subjects from the Ulm-TSST and Reg-TSST (right).

Table 4.1: An overview of each of the three datasets (FAU)- (Reg)- and (Ulm)-TSST used within these initial experiments. Including, number of subjects (#), Age, mean ($\mu$) and standard deviation ($\sigma$) in years - as well as, the speaker independent partitions, Train, (Devel)opment and Test, and the duration (hh: mm) of audio data, before ($\sum$) and after Voice Activity Detection (VAD) and for each TSST task, (Inter)view, and (Arith)methic.

| **Dataset** | **# (M:F)** | **Age $\mu$ / $\sigma$** | **$\sum$** | **VAD** | **Inter.** | **Arith.** | **Train** | **Devel** | **Test** | **$\sum$** |
|---|---|---|---|---|---|---|---|---|---|---|
| FAU | 43 (14:29) | 24.26 / 4.97 | 7:25 | 4:20 | 2:32 | 1:48 | 15 | 15 | 13 | 43 |
| Reg | 27 (13:14) | 22.74 / 2.96 | 4:28 | 2:26 | 1:24 | 1:02 | 10 | 9 | 8 | 27 |
| Ulm | 69 (20:49) | 25.06 / 4.48 | 5:47 | 2:21 | 2:21 | – | 41 | 14 | 14 | 69 |

For both of the experiments the data processing and features used will be the same. All audio was converted to 16 kHz, 16bit, mono, WAV format and normalised to -1 dB before extracting features. A Voice Activity Detection (VAD), utilising the LSTM-RNN approach described by [182] was applied as a first step. From this procedure, in Table 4.1 it can be seen that the arithmetic task contains less speech, and in general, there appears to be substantial silence within the audio data, likely caused by the induced stress. For all datasets segmentation is applied. For FAU-TSST and Reg-TSST, this is based on speech start (provided by the VAD), until the next utterance. As the Ulm-TSST provides transcripts the segmentation is based on this [15].

Each dataset is then partitioned in a speaker-independent manner into training, development, and test sets (see Table 4.1) where demographics including age and gender are balanced as best possible. For this approach a feature-based machine learning method is applied, and the features are extracted with a window size is of 1 second and a hop size of 0.5 seconds. Primarily speech-driven audio features are used, however, video-based features are also included here to observe any advantage that speech may have in this particular context. An overview of the extracted features is shown in Table 4.2.

Table 4.2: An overview of the extracted features, used within these experiments. Of the vision derived features in [15], and for further detail on audio-based features see Chapter 3.

| **Feature Set** | **Modality** | **Dimensions** |
|---|---|---|
| ComParE | Audio | 6 373 |
| eGeMAPS | Audio | 88 |
| DeepSpectrum | Audio | 4 096 |
| VGGish | Audio | 128 |
| FAU-intensity | Video | 17 |
| VGGFace | Video | 512 |

#### 4.1.1.1 Experimental Settings

For all experiments within this section, the tasks are regression in nature, and the same architectures will be applied to model both targets. For an initial data analysis of the cortisol target only, an SVR is first applied. This is then followed by a series of deep learning models based on an LSTM-RNN architecture.

The Support Vector Regression (SVR) algorithm used is an epsilon-support vector regressor with a linear kernel implementation from the Scikit-Learn toolkit [183]. During the development phase for these experiments, a series of SVR models was trained optimising the C parameters ($C \in 10^{-4}$, $10^{-3}$, $10^{-2}$, $10^{-1}$, 1). The model is then re-trained with the concatenated train and development set and evaluated on the test set.

The Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) architecture is based on a similar methodology applied for the baseline of MuSe 2021[1] [32]. In the training processes, the features and labels of every input are further segmented via a windowing approach [184], which may offer the network more context. We experimented with various window lengths, but as in the MuSe Challenge, a window size of 300 steps (150 seconds) was found to be optimal for all datasets. We tested $n = \{1,2,4\}$-layered uni- and bi-directional networks with $h = \{50,100,200\}$ hidden states and a learning rate of $lr = \{0.00005, 0.0001, 0.005, 0.001\}$. Initial experiments showed that the best results were obtained with a 4-layered network, consisting of 2 LSTM and 2 fully-connected (FC) layers, with a hidden size of 50, and a learning rate of 0.00005, (see Figure 4.2). To reduce the computational cost, these parameter values were applied to the experiments reported herein.

For the BPM target a continuous frame-level label is available, which means frame-level predictions using an LSTM-RNN architecture can be obtained and subsequently compared to the target. However, for the cortisol task, only one single target value is available per session at a given time. Moreover, each session lasts approximately ten minutes, and stress may only manifest on short, intermittent segments throughout those recordings. To overcome this, the labels are duplicated on a session-level.

During the training, a many-to-many training is used [185], where the algorithms (SVR and LSTM) are trained to predict the target on all frames. This formulation results in frame-level predictions during evaluation as well. To evaluate the performance of the models, the predictions are first aggregated for each session before comparing them to the reference cortisol values.

The primary evaluation metrics for all models is either Spearman's correlation coefficient $\rho$ or Root Mean Square Error (RMSE) is reported. Correlation as $\rho$ is reported for the cortisol target, as this target is derived from a more ordinal-based sequence. RMSE, in contrast, is

[1] github.com/lstappen/MuSe2021 accessed on: 09.2021

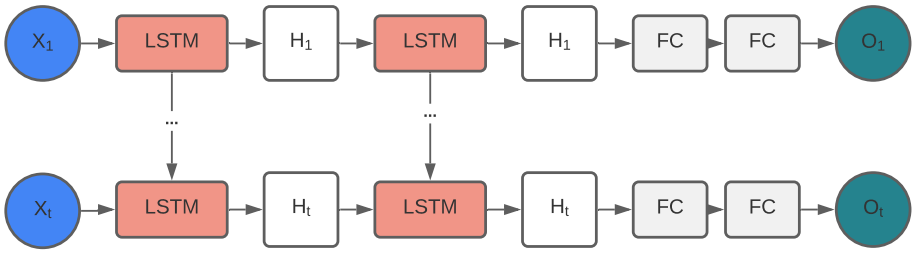



Figure 4.2: LSTM-RNN model architecture. The input sequence $X_i...X_t$ is first fed to two LSTM layers of hidden size 50. The representations of $h_i...h_t$ produced by the second LSTM layer are then processed by two fully-connected layers producing the output sequence $O_i...O_t$.

better suited to a more objective evaluation of the continue signal, which fits the case of continuous heart rate.

### 4.1.2 Sequentially Sampled Cortisol

To evaluate the efficacy of the speech signal as a marker of stress, first, the relationship between sequential cortisol samples and speech-based features is explored. The sequential saliva-based cortisol samples are taken at eight steps, S0 (-45 minutes) to S7 (+60 minutes), measured in nmol/L. The assay (i. e., biochemical analysis procedure) applied to extract cortisol varied for the two datasets in use. FAU-TSST utilised CLIA, and Reg-TSST, DELFIA, meaning that the derived cortisol values are not entirely comparable, for further detail on the difference in these procedures, the interested reader is directed to [114]. For an overview of the raw cortisol in each dataset see the left of Figure 4.1. As can be seen, in Figure 4.1, the behaviour is similar in regards to the time in which the peak of cortisol occurs in both cases at +10 minutes are the stress, with the FAU-TSST dataset sustaining that to +20 minutes. However, as the cortisol of the two datasets is derived with a different assay, and given these statistical differences, the two datasets will be treated individually unless otherwise stated. The primary source of truth for the degree of stress during the TSST setting is the saliva-based cortisol measurements obtained at differing time points. As a traditional cross-corpus analysis would not be fair, the core focus of these experiments is to explore how well the methodology of sequential cortisol prediction can be replicated on the different datasets. However, pooling the data from the two studies and learning a joint model is also explored. Pooling more data, which comes from fundamentally different domains as the two datasets differ in their acoustic nature, might benefit the training procedure particularly in the case of the LSTM-RNN architecture, which would typically require more data to learn from. Thus a model is trained in both single- and multi-domain settings and evaluated on in-domain

data separately for each dataset. The subjects also perform two tasks during the TSST; a spoken interview and an arithmetic task. It is possible that stress manifested differently in the respective acoustic features of these tasks. Thus, the interview and the arithmetic tasks are separated, and a contrast is made to models built after pooling both tasks.

#### 4.1.2.1 Discussion of Results

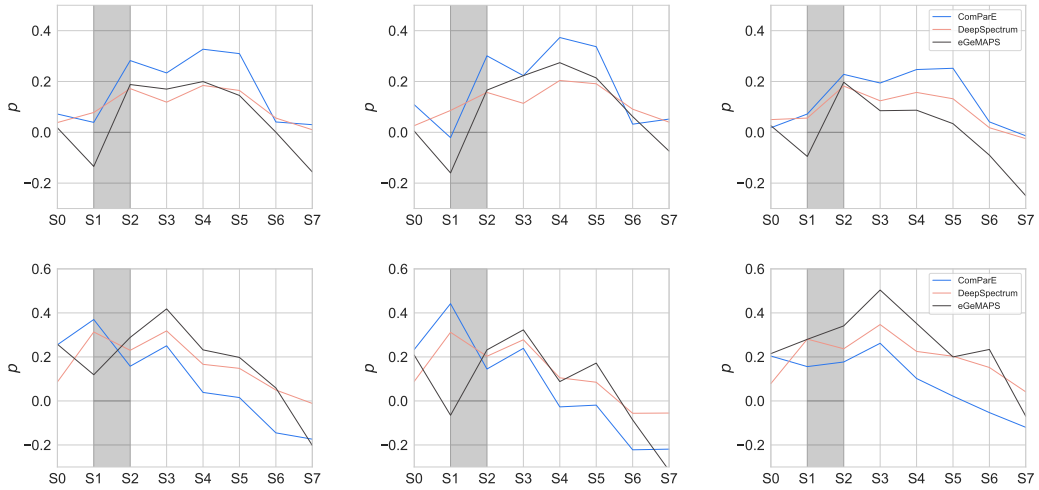



Figure 4.3: The SVR results for the FAU-TSST (above), and Reg-TSST (below). Reporting $\rho$ for all scenarios (left), interview task (middle), and arithmetic task (right).

The conventional SVR and only acoustic features are explored first to observe if the Reg-TSST dataset performs similarly to FAU-TSST in terms of sequential cortisol prediction. In Figure 4.3, the FAU-TSST dataset the correlation is strongest after S3 (+ 10 minutes) S4 (+ 20 minutes). When observing eGeMAPS features this increase is slightly weaker for the arithmetic task compared to the interview, which could be caused by the reduced speech in the arithmetic task. For the Reg-TSST dataset, the trend is less obvious for all feature sets, particularly for the interview task with ComParE features where a strong decline from S1 is seen. The eGeMAPS features appear to perform consistently for both tasks of the Reg-TSST, in this case, the arithmetic task appears to have stronger correlations than the interview however, peaking earlier than FAU-TSST at S3 than, for this task, which may indicate a difference in intra-individual stress response for the two datasets. In general, from these experiments, a higher correlation is obtained post S2 (in most cases S3-S4), suggesting that acoustics features can target stress markers, given the delay common in cortisol response during a period of stress. Hand-crafted features appear to be more suited

Table 4.3: The test set results for session-level cortisol prediction using eGeMAPS features for FAU-TSST, (and multi-domain, plus Reg-TSST)for the (Inter)view and (Arith)metic tasks, as well as the mean ($\mu$) across all and for each individual task. Reporting $\rho$ as the evaluation metric, with values reporting $\geq$.2 positive correlation emphasised.

| $\rho$ | | **FAU-TSST** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Train** | **Task** | **S0** | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** |
| FAU | Inter. | .104 | .016 | .203 | .000 | **.286** | -.209 | -.352 | -.324 |
| FAU | Arith. | **.302** | .060 | **.236** | **.385** | **.396** | -.165 | -.242 | -.225 |
| FAU | Inter. & Arith. | .077 | .093 | .022 | .099 | -.176 | -.286 | -.555 | -.407 |
| FAU & Reg | Inter. | .154 | .055 | -.159 | .159 | .044 | -.341 | .016 | -.456 |
| FAU & Reg | Arith. | **.335** | **.214** | **.368** | **.374** | **.698** | **.286** | -.027 | -.214 |
| FAU & Reg | Inter. & Arith. | .126 | **.209** | -.077 | .104 | .088 | -.220 | -.632 | -.456 |
| | Inter $\mu$ | .129 | .035 | .022 | .159 | .165 | -.275 | -.168 | -.390 |
| FAU | Arith $\mu$ | .318 | .137 | **.302** | **.379** | **.547** | .061 | -.135 | -.220 |
| | All $\mu$ | .183 | .108. | .099 | .187 | **.223** | -.156 | -.299 | -.347 |

for this task overall, and given this, eGeMAPS will be used as the main acoustic feature set for all further experiments.

The LSTM-RNN model results are shown in Table 4.3 for FAU-TSST evaluated experiments and Table 4.4 for Reg-TSST evaluation. Again, speech-based models can predict cortisol levels samples taken at time points S2-S5 with a medium to strong correlation and a mean peak around S4 (+20 minutes after the TSST) in the case of FAU-TSST. This is consistent across both datasets and tasks. However, there are important and interesting differences across different settings. In general, the LSTM-RNN can better predict cortisol from the arithmetic task of FAU-TSST, which slightly contradicts the SVR results and shows that this task can also yield good results if the sequential nature of different frames is considered. This indicates that, for this dataset, subjects either became more stressed during the arithmetic part of the TSST or that the manifestation of stress in the speech was more pronounced.

Overall, for both datasets, a higher correlation for point S3-S4 is seen, with the interview task tending to peak a bit earlier than the arithmetic one. Given the relative delay between the two tasks, this is in line with the authors previous research [112] showing that speech signals are more correlated with cortisol measurements taken approximately 10 minutes after initial stress. Interestingly, a high correlation is occasionally seen for cortisol measures taken at S1 (1 minute *before* the TSST) for Reg-TSST (particularly for the interview task). which seems counter-intuitive, however, it could be considered that this is attributed to the lower variability across subjects for measurements at S1 (see Figure 4.1), which may have made this task easier to learn.

Finally, the multi-domain models built by pooling both datasets perform consistently better while additionally benefiting from the pooling of the interview and arithmetic tasks in the case of Reg-TSST. This illustrates that, even though the cortisol measurements in the two datasets are based on fundamentally different scales, the relationship between relative cortisol values and acoustic features remains consistent, allowing the models to benefit from more diverse data and obtain better performance, as measured by $\rho$ correlation.

Table 4.4: The test set results for session-level cortisol prediction using eGeMAPS features for the Reg-TSST (and multi-domain, plus FAU-TSST) for the (Inter)view and (Arith)metic tasks, as well as the mean ($\mu$) across all and for each task. Reporting $\rho$ as the evaluation metric, with values reporting $\geq$0.2 positive correlation emphasised.

| $\rho$ | | **Reg-TSST** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Train** | **Task** | **S0** | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** |
| Reg | Inter. | **.297** | **.827** | **.527** | **.261** | **.236** | -.127 | -.527 | -.079 |
| Reg | Arith. | .091 | .559 | -.164 | .091 | **.455** | **.333** | **.552** | **.406** |
| Reg | Inter. & Arith. | .127 | **.474** | .055 | **.285** | **.248** | .115 | -.273 | -.406 |
| FAU & Reg | Inter. | -.152 | **.559** | **.467** | **.200** | **.261** | -.018 | -.539 | .164 |
| FAU & Reg | Arith. | -.212 | **.267** | .055 | -.042 | **.370** | **.212** | .188 | .091 |
| FAU & Reg | Inter. & Arith. | .006 | **.584** | **.721** | **.770** | **.442** | .176 | -.139 | -.042 |
| | Inter $\mu$ | .072 | **.693** | **.497** | **.230** | **.262** | -.073 | -.533 | .043 |
| Reg | Arith $\mu$ | -.061 | **.413** | -.055 | .025 | **.412** | **.273** | **.370** | **.285** |
| | All $\mu$ | .026 | **.545** | **.279** | **.261** | **.335** | .115 | -.123 | .022 |

To compare the performance of audio, video-based models for stress recognition on the FAU-TSST dataset were also trained. Using an identical experimental protocol, and simply substituting eGeMAPS with VGGFace (VGGFace) features. Results are shown in Figure 4.4, and as can be seen, the vision features are in general lower than those obtained with eGeMAPS features. This indicates that, particularly in the FAU-TSST dataset, the auditory modality is more appropriate as a marker of stress.

Moreover, both early and late multimodal fusion is applied for these experiments. For early fusion, the features are concatenated and a single model is trained, for the decision-level (late fusion) separate models are trained individually for each modality, and the predictions of these are fused by training an additional uni-directional LSTM-RNN model as described above. From these it is observed that multimodal fusion can lead to better performance in some cases, most notably for the prediction of cortisol at S2, suggesting that the interview task was more meaningful for these features – potentially due to more facial activity. However, generally eGeMAPS features remain strong as a uni-modal approach.

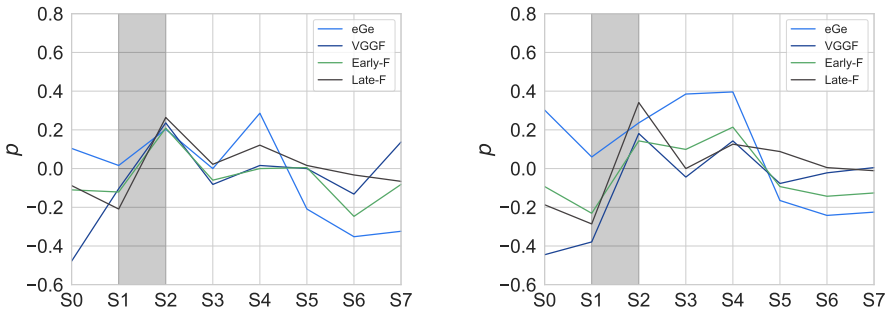



Figure 4.4: The correlation $\rho$ for session-level cortisol prediction using VGGFace and eGeMAPS features on the data from FAU-TSST interview task (left) and arithmetic task (right). Reporting uni-modal visual-based results as well as multi-modal fusion utilising either an early or late fusion a strategy.

### 4.1.3 Continuous Heart Rate and Cortisol

Next, in an endeavour to evaluate the ability of the speech signal to model markers of stress, in a continuous manner, heart rate is evaluated in relation to cortisol. Stress is known to impact Heart Rate (HR) [186, 187] through its activation of the sympathetic [188] and suppression of the parasympathetic branch of the autonomic nervous system [189]. HR itself can therefore serve as a vital indicator of stress in modern affective computing applications. As discussed earlier, only one of the three datasets utilised, the Reg-TSST dataset, has both HR and cortisol measurements, whereas the FAU-TSST dataset has only cortisol measures and Ulm-TSST only HR. Thus, the only dataset where the relationship of HR with stress (as cortisol as ground truth) can be evaluated fully is Reg-TSST.

#### 4.1.3.1 Discussion of Results

To obtain an understanding of the HR signal in relation to cortisol, Figure 4.5 shows the distribution of ground truth HR values for the Reg-TSST dataset in groupings of Low (below the 33rd percentile), Mid, and High (above the 66th percentile) cortisol levels taken at different time points. Due to the differing assay between FAU-TSST and Reg-TSST, when discussing these groupings, they are based on the percentile distribution of the cortisol samples, which for FAU-TSST is 1) 33rd $< 4.90$ nmol/L, 2) middle $4.90 - 9.05$ nmol/L, 3) 66th $> 9.05$ nmol/L , and for Reg-TSST 1) 33rd $< 4.18$ nmol/L, 2) middle $4.18 - 6.79$ nmol/L, 3) 66th $> 6.79$ nmol/L .

When observing all of the results for each sequential time step, a two-sample independent T-tests shows that all results are significant at the $p < 0.05$ level, except the low vs high percentiles at time S0 and the low vs middle percentiles at time S5. Overall, a rising trend for

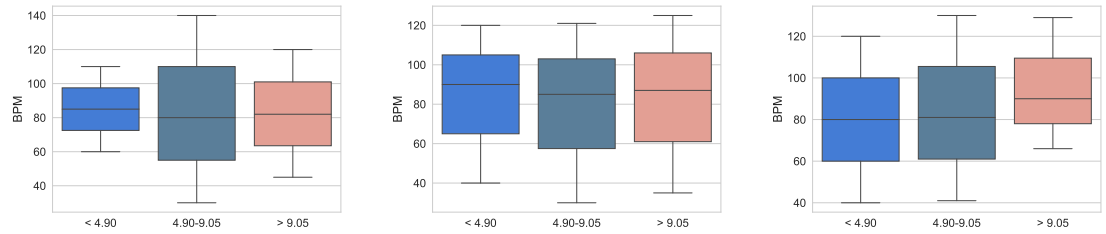



Figure 4.5: Box plots of ground truth BPM values for the Reg-TSST dataset. Grouped based on raw cortisol (nmol /L) measures taken at time-points S2,S3,S5.

HR as BPM shows an increase in cortisol levels; this is consistent with our expectations and prior work [186, 187]. This trend is particularly pronounced for S5 (+20 minutes after the TSST), showing that higher cortisol values obtained during that time were highly correlated with higher HR during the TSST.

The other dataset used in this study with cortisol measurements is FAU-TSST, but this dataset does not have available HR measures, and so a model built on the other two datasets can be applied to predict HR as BPM. For this, the speech modality of the Reg-TSST and Ulm-TSST datasets are used to build a model, which can then predict BPM on the FAU-TSST dataset. This is motivated by audios commonality across the three datasets, and also the effect of HR on the voice has long been established by previous research [190]. Several prior works have attempted to model HR from voice signals, either as a classification [191] or a regression task [192]. In [192] the authors use eGeMAPS to predict BPM from speech on the segment level, and achieve an RMSE of 12 BPM. Inspired by these past findings, HR in the form of BPM is targeted using the acoustic speech-based features. As all three datasets were recorded in different locations with potentially different acoustic conditions, the domain mismatch problem may be a concern [193], where models trained on data from one domain might not generalise well to different domains. Moreover, the two datasets are slightly different for their range in BPM ranges, with subjects in Reg-TSST having a generally lower BPM than subjects in Ulm-TSST. To address this issue, two single-domain models are trained using both datasets in isolation and then trained in a multi-domain model using data from both datasets together. In all cases, the performance is evaluated and reported separately for each dataset.

The RMSE results are shown in Table 4.5. The initial observation shows that all models perform better on the Ulm-TSST dataset and that in-domain models perform better than their cross-domain counterparts. Moreover, the multi-domain model does not bring any improvements compared to the single-domain ones. The limited overlap in the BPM ranges for the two datasets is most likely the reason for this; combining the data does not lead

Table 4.5: The (devel)opment and test RMSE results for BPM prediction in a single- and multi-domain paradigm. Utilising the Reg-TSST and Ulm-TSST datasets with the eGeMAPS and an LSTM-RNN architecture.

| RMSE | **Reg-TSST** | | **Ulm-TSST** | |
|---|---|---|---|---|
| **Train** | **Devel** | **Test** | **Devel** | **Test** |
| Reg-TSST | **39.90** | **38.57** | 20.98 | 22.96 |
| Ulm-TSST | 36.53 | 40.80 | **19.32** | **22.70** |
| Reg-TSST & Ulm-TSST | 36.23 | 38.96 | 23.07 | 23.05 |

to considerable benefits since the target is different. The best performing combination is obtained when training and testing on the Ulm-TSST dataset and achieves an RMSE of 19 BPM, which is not as strong as the previously mentioned state-of-the-art of 12 BPM [192]. It is worth noting that capturing biological signals is a challenge, with the potential movements of the subjects leading to more unreliable measurements, which in turn makes the target much more of a challenge to learn.

Despite the relatively low performance obtained by these speech-to-BPM models, they can still be used to obtain BPM predictions on the FAU-TSST dataset, as the primary interest is in the usefulness of predicted BPM values for stress modelling. In Figure 4.6, the distribution of predicted BPM values for cortisol measurements obtained at different time points are shown. A slight downward trend for BPM is observed as the cortisol level increases in this case which would be different to the trend seen in Figure 4.5. However, rather than these low measurements implying that stress leads to a lower BPM, they can be interpreted as a demonstration that BPM signals, though theoretically well justified as predictors of stress, are nevertheless a challenge to collect in practice. Thus, BPM alone may be inferior to signals like voice that are easier to manage and provide richer information for evaluation. Although the trend is not what would be expected, there is still a separation between different cortisol levels, indicating that predicting HR from the speech signal can be a valuable proxy for stress prediction. Two-sample independent t-tests show that all differences are significant at the $p < 0.05$ level except the middle vs high percentiles as measured at S4.

### 4.1.4 Conclusions

When utilising acoustic information to recognise markers of stress, it can be seen from these experiments that audio is a robust modality, for targeting cortisol as compared to HR and other vision-based features (**RQ-1**). In this way, speech, in general, does show to be a valuable proxy for the degree of stress concerning percentile groupings of cortisol, and

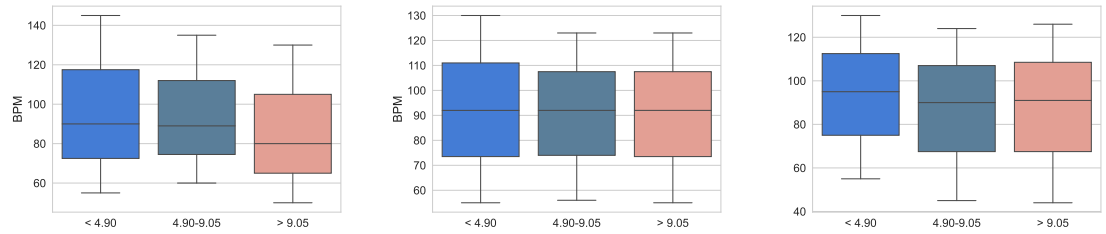



Figure 4.6: Box plots of predicted BPM values for the FAU-TSST dataset, based on a combined model train on the FAU-TSST and Ulm-TSST datasets. Grouped based on raw cortisol (nmol/L) measures taken at time-points S2,S3,S5.

markers of stress in general. The main results from these experiments are the validity of speech-based prediction of high-levels of cortisol, which is substantiated by results which show that saliva-samples taken between 10–20 minutes after the stress event are more highly correlated to the speech features in general. This finding links to literature that supports the delay in cortisol after stress (in this case, when speaking).

From a uni-modal vs multimodal perspective (**RQ-2**), the system benefited when audio and video-derived features were fused, which may be partly due to the sustained silence that exists for some more heavily stressed subjects. However, as a uni-modal signal, handcrafted audio features appear to model the targets well on their own, and when fusing modalities, there did not appear to be a substantial benefit to neither early nor late fusion, although late fusion showed a slight increase in overall correlation. In general, this study was also a strong example of interdisciplinary work (**RQ-4**). The data collection was made solely by the respective psychology focused research groups at the University of the Regensburg, the University of Ulm, and the University of FAU Erlangen-Nürnberg, which not only gives the data more scientific validity, this also means that via interdisciplinary collaboration with the data collectors, several aspects of the experimental settings can be rigorously considered, e. g., the effect of the varied assay used for extracting cortisol **(RQ-3)**.

In summary, these experiments show that audio alone is suitable for recognising several physiological markers of stress; however, a multimodal approach is beneficial where speech is limited. Despite this, as with many states of emotional wellbeing, there appears to be substantial variance in the physiological manifestation of stress, making generalisation a challenge, particularly in the case of BPM. This strong variance may be tackled by personalised machine learning approaches, and should be considered in future work.

## 4.2 The State of Anxiety in Speech

The rate of diagnosis for mental health disorders characterised as anxiety disorders have been increasing throughout the last decade, particularly for those living in urban environments [194]. There are several anxiety disorders that range in their severity, including Generalised Anxiety Disorder (GAD), Obsessive-Compulsive Disorder (OCD), and Post-Traumatic Stress Disorder (PTSD). For the current experiment, GAD (henceforth, anxiety) is the focus and has been defined as excessive worry and apprehension occurring more days than not [195]. A feeling of uncertainty is often a catalyst for anxiety, and the current global pandemic of SARS-CoV-2 now contributes to this [196]. With this in mind, as with many states of emotional wellbeing, mechanisms to monitor and treat anxiety are needed [197].

The proceeding experiments are based on those initially described in [19]. Within these experiments the effect of anxiety on speech is explored, via an acoustic analysis, and machine learning experiments focusing on the ability for the speech signal to computationally predict the degree of anxiety (based on the BAI score). The experiments focus on four classes of emotional sustained vowels (*Sad*, *Smiling*, *Comfortable*, and *Powerful*) which have been taken from the Düsseldorf Anxiety Corpus (DAC) as described with more detail in Chapter 3. First, acoustic analysis is performed on the data, observing a selection of conventional LLDs, concerning BAI groupings. Following this, a series of regression experiments are performed to see how well acoustic features can target the BAI score within particular BAI groupings and with consideration to certain questions within the BAI that may be particularly related to vocalisations. As with the previously described study, these experiments address several RQs in the following manner:

- **RQ-1:** Validating the use of audio-based features for targeting states of reduced emotional wellbeing. This is more deeply explored with these experiments as a specific focus on prosodic-derives features of the speech signal are analysed, as well as targeting a well established ground-truth for anxiety.
- **RQ-3:** Utilising the DAC dataset to explore the available data in the community, which is of a larger scale than many other datasets. Further to this, for these experiments a self-reported measure is used, and so the validity of targeting this in comparison to other metrics, e. g., perceived or other objective markers, from audio will be discussed.

### 4.2.1 Data and Procedure

As a first step, the audio was converted to 16 kHz, 16 bit, mono, WAV format, and as the beginning and end of many instances contained silence, the Librosa toolkit trimming function

Table 4.6: The speaker (#) independent partitions, Train, (Devel)opment, and Test. Gender (M)ale:(F)emale, and number of (inst)ances.

| | **Train** | **Devel** | **Test** | $\sum$ |
|---|---|---|---|---|
| # | 74 | 97 | 68 | 239 |
| M:F | 26:48 | 25:72 | 19:49 | 69:170 |
| Inst. | 614 | 511 | 440 | 1565 |

was applied to automatically trim each file. Given the nature of the sustained vowels being used this was trivial, however the data loss was quite substantial, with the original data duration (4 h:30 :m24 s) reduced to 3 h:00 m:40 s. Proceeding this, from the 239 speakers (69 males) which are taken from the DAC, speaker-independent partitions are created, train, development, and test (see Table 4.6).

Within the dataset, the absolute BAI rating for a given individual range from 0–60 and to avoid weighting for particular speakers, these raw annotations were standardised to zero mean and unit standard deviation, resulting in a range of [-1.11: 4.36]. Speakers were grouped based on the raw BAI values, as either High-BAI or Low-BAI ($\leq$ 20: Low, $\geq$ 21: High).

The procedure for the BAI is a series of questions that relate to the degree (Likert scale 0–3) to which a given symptom bothers them in the past month. To explore this more specifically for speech, there are two questions which relate more specifically to vocalisation *feeling of choking* (choking) and a *difficulty breathing* (breathing), and so these are used to group the data in another manner. Where no symptoms (no) would be those reporting a value for 0 (Not at all) has symptoms (has) individual reporting 1 and above (Mildly, Moderately, Severely) for each of those questions.

The experiments being conducted are based on a feature-based machine learning strategy, and the features used include both hand-crafted and spectrogram representations of speech (see Table 4.7). There was no window or hop applied in this case as the duration for the sample is reasonably small (ca. 8 seconds), and there is no words to assume a meaningful segmentation on, and so extraction was made over the entire signal. No other modalities are extracted from the data in this case, as this was not made available in the DAC.

#### 4.2.1.1 Experimental Settings

For the machine learning experiments where the degree of anxiety is targeted by the acoustic features, given that the subset being used from DAC is reasonably small (ca. 3 hours), for a robust and easily reproducible approach, an epsilon- SVR with a linear kernel is applied. During the development phase, a series of SVR models are trained, optimising the complexity

Table 4.7: An overview of the extracted features used within these experiments, for further detail on these feature sets see Chapter 3.

| **Feature Set** | **Modality** | **Dimensions** |
|---|---|---|
| ComParE | Audio | 6 373 |
| eGeMAPS | Audio | 88 |
| DeepSpectrum | Audio | 4 096 |

parameters ($C \in 10^{-4} - 1$), and evaluating the performance on the development set. The best model is then re-trained with the concatenated train and development set and evaluated on the test set. This is then repeated for each combination.

To evaluate the results of the machine learning experiments, the Spearman's correlation coefficient ($\rho$) is applied, due to the ordinal nature of the raw BAI values. Additionally, Cohen's $d$ is reported as a measure of effect size between the various groupings, and for the machine learning experiments, this proceeds an evaluation of each test set prediction result for normality using a Shapiro-Wilktest [198], as well as two-tailed T-test, rejecting the null hypothesis at a significance level of $p < 0.05$. In general the effect size can also be interpreted as, small $d \geq .2$, medium $d \geq .4$, and large $d \geq .8$.

### 4.2.2 Acoustic Analysis of Anxious Speech

There has been limited analysis in the Computational Paralinguistics (CP) and speech processing community for the state of anxiety. In [199] the authors compare self-reported anxiety from the renowned State-Trait Anxiety Inventory (STAI) [200], to human perception and compare this with acoustic LLDs, yet there appears to be no literature that does this in the context of the BAI score. Inspired by results in [199], based on the groupings created earlier, a variety of acoustic LLDs will be extracted from each of the speech samples and analysed.

#### 4.2.2.1 Discussion of Results

As a brief initial step, the effect size (Cohen's $d$) between High-BAI and Low-BAI groupings of each of the acoustic feature sets extracted for each sustained vowel is evaluated (see Table 4.8). Of note from this analysis, it can be seen that DeepSpectrum features appear to have consistently moderate effect sizes, larger than ComParE and eGeMAPS, particularly for the *Sad* and *Comfortable* class. However, in most cases there is a difference in the standard deviation of the mean from all acoustic feature sets, taken from all sustained vowels.

Table 4.8: The effect size as Cohen's $d$, between the mean of all features sets (eGeMAPS, DeepSpectrum, and ComParE for Low-BAI vs High-BAI groupings of each stressed vowel class. Individual results excluded reject the null-hypothesis.

| **Feature Set** | **Sad** | **Smile** | **Comf.** | **Power** |
|---|---|---|---|---|
| eGeMAPS | .564 | .463 | -.053 | .649 |
| ComParE | .774 | .336 | -1.339 | – |
| DeepSpectrum | .819 | .708 | .797 | .463 |
| $\mu$ | .719 | .502 | -.198 | .556 |

Table 4.9: The results from the acoustic analysis. Reporting the mean across all speakers for the mean ($\mu$) and standard deviation ($\sigma$) of $F_0$ (Hz), HNR (dB), and (Int)ensity (dB), for each class of emotionally sustained vowel in (Low) and (High)-BAI groupings. Including the number (#) of samples in each group.

| **BAI-level** (#) | **Class** | $F_0$ $\mu$ | $F_0$ $\sigma$ | **HNR** $\mu$ | **HNR** $\sigma$ | **Int.** $\mu$ | **Int.** $\sigma$ |
|---|---|---|---|---|---|---|---|
| Low (290) | *Sad* | 165.23 | 24.73 | 15.93 | 3.86 | 54.65 | 9.80 |
| High (88) | | 148.04 | 27.093 | 13.97 | 4.00 | 56.68 | 9.39 |
| Low (302) | *Smile* | 221.88 | 25.09 | 18.07 | 4.03 | 60.66 | 11.87 |
| High (89) | | 185.09 | 20.85 | 16.31 | 3.80 | 62.42 | 10.95 |
| Low (305) | *Comf.* | 171.18 | 16.10 | 16.66 | 3.41 | 56.42 | 11.64 |
| High (88) | | 155.65 | 22.03 | 15.45 | 3.71 | 60.23 | 11.33 |
| Low (93) | *Power* | 193.87 | 12.07 | 20.16 | 3.74 | 64.75 | 13.94 |
| High (310) | | 170.08 | 13.98 | 17.59 | 3.73 | 65.75 | 12.56 |

Next, the standard deviation (STD) and the mean of Pitch $F_0$ (Hz), intensity (dB), and HNR (dB) for each speech sample is extracted for the subset of the DAC, and also the Low-BAI and High-BAI pairings, for which the effect size using Cohen's $d$, will be compared between. Before this, a two-tailed T-test is performed, rejecting the null hypothesis at a significance level of $p < 0.05$. An overview of the mean results is given in Table 4.9.

When evaluating pitch ($F_0$) of the four classes of sustained vowels, higher STD between Low-BAI and High-BAI groupings for all classes is found, except *Smiling* – particularly, for *Sad* and Comfortable, which show a smaller and medium effect size, respectively. This finding shows that lower aroused phonation types present stronger $F_0$ variance for individuals with higher anxiety levels. In Figure 4.7 this finding is shown more specifically for a selection of speakers. Where those with higher levels of BAI have a lower STD for $F_0$ for *Smiling* , but higher for Sad, with this finding seemingly consistent between genders.

For the intensity of the speech signal, in all cases samples of Low-BAI show strong deviation in dB, and particularly for *Smiling* and *Powerful*. Additionally, when comparing

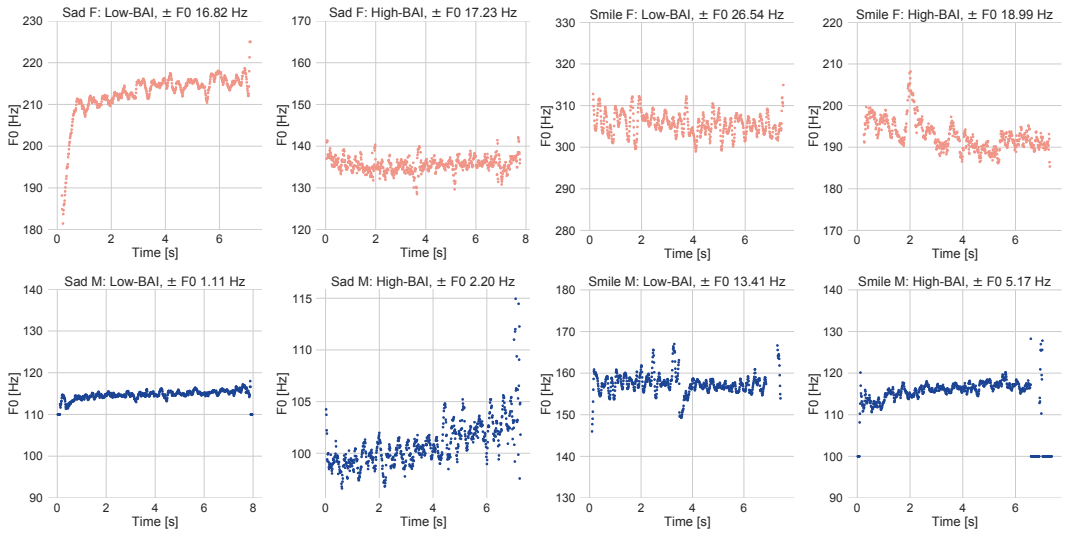



Figure 4.7: Pitch $F_0$ plot for(F)emale and (M)ale samples from the Low-BAI and High-BAI groupings, taken from the *Sad* and *Smile* sustained vowel. Of note a higher mean standard deviation of $F_0$ for samples in High-BAI grouping as well as for females compared to males is shown. Further to this, the $F_0$ for High-BAI grouping also appears to be in a lower range as compared to the Low-BAI for each gender.

Low-BAI-*Sad* and Low-BAI-Powerful, a large effect size ($d$=1.068) is found, reaffirming the difference in affect for the target emotional style for these sustained vowels. Overall the mean intensity is quite consistent with a general increase from *Sad* to *Powerful*.

For HNR, all classes show a higher mean results for the Low-BAI class but a higher STD in most cases for the HNR for the High-BAI, aside from the *Smiling* vocalisation. The STD finding is more substantial significant for *Sad* and *Comfortable* and shows that vocal-fold action is less consistent for these classes in the High-BAI group.

The findings from the initial acoustic feature set analysis, specifically for DeepSpectrum, also seems to be reflective of the individual LLDs, specifically for *Sad* and *Comfortable*. This leads to the assumption that given the visual nature of DeepSpectrum features, an increased STD in $F_0$ for those with higher anxiety may be more easily captured with these features. Further to this, DeepSpectrum most likely observes noise in the signal, as reflected by the varied HNR for both *Sad* and *Comfortable*. From this analysis the relation between state of emotional arousal for the sustained vowel and the $F_0$ STD is clearly shown.

### 4.2.3 Anxiety Prediction from Stressed Vowels

A series of experiments are performed to explore the efficacy of computational prediction of anxiety from speech in a set domain-specific groupings, in relation the BAI scale. The

Table 4.10: The results for the prediction of Beck Anxiety Inventory (BAI) for *Sad* and *Smiling* sustained vowels and (Has) symptoms or (No) symptoms of *feeling of choking* and *difficulty in breathing*. Reporting $\rho$, and for *Has* symptoms $*$ indicates significance ($p < 0.05$) compared to the equivalent *No* symptoms. Emphasised test results show $\rho > .3$.

| $\rho$ | **Sad** | | | | **Smiling** | | | |
|---|---|---|---|---|---|---|---|---|
| **Feeling of choking** | No | | Has | | No | | Has | |
| | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | .102 | .190 | .611 | -.064* | .252 | **.350** | .343 | .218* |
| ComParE | .099 | -.170 | .110 | .051* | .309 | .011 | .223 | **.535*** |
| DeepSpectrum | .078 | .288 | .369 | -.397* | .202 | .246 | .022 | -.160* |
| **Difficulty in breathing** | No | | Has | | No | | Has | |
| | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | -.019 | .120 | .276 | -.122* | .187 | .255 | .453 | **.340** |
| ComParE | .048 | -.139 | -.028 | **.363*** | .282 | -.021 | .357 | **.699*** |
| DeepSpectrum | -.006 | **.357** | .301 | .284* | .342 | .256 | -.045 | .028* |

same subsets as applied earlier are used for these experiments, as well as the four emotional sustained vowel classes, described previously (*Sad*, *Smiling*, *Comfortable*, and *Powerful*). Experiments are performed from within the emotion-class, as well as for all together.

#### 4.2.3.1 Discussion of Results

The results for experiments are divided into more manageable groups for east of discussion In Table 4.11, Table 4.12 and Table 4.13 correlation prediction results for BAI of *Low* and High-BAI groupings are given for the *Sad* and *Smiling*, *Comfortable* and *Powerful*, and All sustained vowels respectively. For the *Has* and *No-symptoms* groupings, Table 4.10, Table 4.14, and Table 4.15 report results for the *Sad* and *Smiling*, *Comfortable* and Powerful, and All sustained vowels, respectively.

In general, as indicated by $*$, there are significant differences in almost all predictions for Low-BAI vs High-BAI groupings. As well as this, in most cases, High-BAI grouped results are substantially higher than Low-BAI grouped results. Although the results vary slightly, they do suggest that speech characteristics, harnessed for the prediction of anxiety, are more robust when anxiety is at high levels. This finding is supported by earlier discussed literature, which suggests that speech disturbances and varied speech rate are prominent in the speech of those with high anxiety [201].

Looking closer at the BAI grouped experiments, High-BAI grouped anxiety predictions, in general, are stronger, with at best, .506 $\rho$ for prediction of BAI across all sustained vowels (see Table 4.13). Through the late-fusion of the two best results eGeMAPS and DeepSpectrum, this is increased to .592 $\rho$. For the individual sustained vowels (see Table 4.11), *Smiling* in High-BAI grouping performs best, with eGeMAPS showing up to .593 $\rho$, a

Table 4.11: The results for the prediction of Beck Anxiety Inventory (BAI) for *Sad* and *Smiling* sustained vowels and Low- and *High-BAI* groupings. Reporting $\rho$, and for High-BAI $*$ indicates significance ($p < 0.05$) compared to the equivalent *Low-BAI*. Late fusion is included from the mean of predictions of the two best performing feature sets. Emphasised test results show $\rho > .3$.

| $\rho$ | **Sad** | | | | **Smiling** | | | |
|---|---|---|---|---|---|---|---|---|
| **BAI** | Low | | High | | Low | | High | |
| | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | -.049 | .106 | .210 | -.057* | .043 | -.012 | .181 | **.593*** |
| ComParE | .025 | -.018 | .100 | -.241* | .015 | -.271 | .441 | **.446*** |
| DeepSpectrum | .104 | .210 | .005 | **.304*** | .190 | .012 | .253 | **.418*** |
| **Late-Fusion** | – | .194 | – | **.228*** | – | .008 | – | **.646*** |

Table 4.12: The results for the prediction of Beck Anxiety Inventory (BAI) for *Comfortable* and *Powerful* sustained vowels and Low and High BAI grouping. Reporting $\rho$, and for *High-BAI* $*$ indicates significance ($p < 0.05$) compared to the equivalent *Low-BAI*. Late fusion is included from the mean of predictions of the two best performing feature sets.

| $\rho$ | **Comfortable** | | | | **Powerful** | | | |
|---|---|---|---|---|---|---|---|---|
| **BAI** | Low | | High | | Low | | High | |
| | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | -.053 | .029 | .350 | .031* | -.098 | -.294 | .084 | .087* |
| ComParE | -.101 | .183 | .154 | -.127* | .096 | .073 | .252 | .143* |
| DeepSpectrum | .219 | .216 | .146 | -.145* | .132 | -.004 | .146 | -.501* |
| **Late-Fusion** | – | .213 | – | .027* | – | -.029 | – | .167 |

result which is also improved by late-fusion up to .646 $\rho$. A slight moderate correlation for DeepSpectrum of *Sad* High-BAI grouping is seen. However, this is not consistent with all feature sets. For *Comfortable* and *Powerful*, there are no substantial correlations overall. When comparing to *Smiling* and *Sad* this leads to the assumption that *Comfortable* and *Powerful* do not provide meaningful information for the current task, and that the inherent emotionality of the *Sad* and *Smiling* is very meaningful in this context.

In regards to the grouping of *Has-symptoms*, or *No-symptoms* of *feeling of choking*, *Smiling* samples again performs best, with ComParE at best .535 $\rho$. However, in this case, eGeMAPS and DeepSpectrum are less able to capture the phenomena. *Comfortable* phonations show a strong negative correlation for the *Has-symptom choking* grouping, a finding which to a degree also appears for *Sad*, suggesting (based on acoustic analysis) that intensity may play a strong roll in this task. When predicting BAI from all samples with *No-symptom* of *choking*, this appears to be stronger than the *Has-symptoms* pairing. Overall, there are no strong findings from this paradigm. However, most *No-symptoms*

Table 4.13: The results for the prediction of Beck Anxiety Inventory (BAI) for all Düsseldorf Anxiety Corpus (DAC) data with the same Low vs High-BAI groupings.

| $\rho$ | **All** | | | |
|---|---|---|---|---|
| **BAI** | Low | | High | |
| | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | .012 | -.033 | .173 | **.405*** |
| ComParE | .077 | .002 | .120 | .120* |
| DeepSpectrum | .141 | .286 | .105 | **.506*** |
| **Late-Fusion** | – | .238 | – | **.592*** |

Table 4.14: The results for the prediction of Beck Anxiety Inventory (BAI) for *Comfortable* and *Powerful* sustained vowels and (Has) symptoms or (No) symptoms of *feeling of choking* and *difficulty in breathing*. Reporting $\rho$, and for *Has* symptoms $*$ indicates significance ($p < 0.05$) compared to the equivalent *No* symptoms. Emphasised test results show $\rho > .3$.

| $\rho$ | **Comfortable** | | | | **Powerful** | | | |
|---|---|---|---|---|---|---|---|---|
| **Feeling of choking** | No | | Has | | No | | Has | |
| | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | .192 | .067 | .132 | -.424 | .039 | **.376** | -.148 | **.317*** |
| ComParE | -.008 | .294 | .081 | -.463 | .057 | .201 | .392 | **.494*** |
| DeepSpectrum | .130 | .300 | -.003 | .297* | .106 | **.438** | .706 | .200* |
| **Difficulty in breathing** | No | | Has | | No | | Has | |
| | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | .036 | .067 | .413 | -.217 | .081 | .234 | .218 | .194* |
| ComParE | .090 | .284 | .082 | .078* | .099 | .212 | .259 | -.384* |
| DeepSpectrum | .004 | **.302** | -.010 | .169* | .176 | **.384** | -.009 | -.026* |

grouped results perform better than *Has-symptoms* grouped, which suggests a need for further acoustic analysis to observe any variation in the samples for this constellation.

For the grouping of *Has-symptoms* or *No-symptoms* of *difficulty in breathing*, it is seen, as with *choking*, that the *No-symptoms* grouped results are often stronger than *Has-symptoms* group. However, across feature sets, this is somewhat confused. For *Sad*, for example, the *No-symptoms* grouping performs better with DeepSpectrum, but overall, ComParE shows slightly better results for the *Has-symptoms* grouping. Like all other groupings, the *Smiling* class in the *Has-symptoms* grouping shows the best result, up to .699 $\rho$. ComParE also performs best when utilising all data for the *Has-symptoms* grouping. This is suggesting that HNR, which may be stronger due to restricted airflow, is more easily captured by ComParE features for individuals with this *breath* symptom.

To evaluate the degree to which highly anxious speech can be applied to predict an individuals BAI further, the experiment was rerun with all data and without any groupings, i. e., without Low vs High-BAI, or Has vs No symptoms (see Table 4.16). From this, the

Table 4.15: The results for the prediction of Beck Anxiety Inventory (BAI) for all Düsseldorf Anxiety Corpus (DAC) data, utilising the *Has* vs *No* symptoms groupings.

| $\rho$ | **All** | | | |
|---|---|---|---|---|
| **Feeling of Choking** | No | | Has | |
| | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | .146 | .222 | .103 | -.029* |
| ComParE | .102 | -.163 | -.148 | -.392* |
| DeepSpectrum | .188 | .254 | .075 | .118* |
| **Difficulty in Breathing** | No | | Has | |
| | **Devel** | **Test** | **Devel** | **Test** |
| eGeMAPS | -.019 | .120 | .187 | .255* |
| ComParE | .112 | .136 | .237 | **.379*** |
| DeepSpectrum | .151 | .285 | .178 | .126* |

Table 4.16: The results for prediction of Beck Anxiety Inventory (BAI) from all Düsseldorf Anxiety Corpus (DAC) data combined, without groupings. Late-fusion was calculated based on the mean from the two best performing feature sets.

| $\rho$ | **Devel** | **Test** |
|---|---|---|
| eGeMAPS | .189 | .245 |
| ComParE | .093 | .213 |
| DeepSpectrum | .106 | .238 |
| **Late-fusion** | – | **.243** |

original High-BAI grouped results with DeepSpectrum and eGeMAPS remain stronger, with the best results from late fusion being .243 $\rho$ for all data (a result which can be considered negligible correlation). This result is significantly lower than the best High-BAI result with all sustained vowels, reporting a very large effect size ($d$=1.718) when comparing the two best results. In general, for all scenarios, the *Smiling* class performs best for the stronger High-BAI and *Has-symptoms* groupings. This finding could suggest that anxiety is more prevalent in a more facially strained stressed vowel. There is much in the literature relating to *Smiling* and anxiety, for example, the "fooled by a smile" effect in which those who suffer from anxiety can show untrue emotional expressions [94]. Furthermore, high anxiety involves much more facial expression and general movement, as compared to lower anxiety, with 'non-enjoyment' smiles being displayed frequently [202].

### 4.2.4 Conclusions

In this section, the effect of anxiety on speech was evaluated, and the efficacy of predicting indications of anxiety i. e., the BAI score, from non-lexical sustained vowels were evaluated. The findings show that utilising speech-based features for the prediction of anxiety is valid

and that recognition of higher levels of anxiety is more easily targeted by acoustic features than lower levels of anxiety (**RQ-1**). As individuals reporting high levels of BAI may need a more timely intervention, this finding is promising and particularly applicable to integration with a more empathic AI.

Similarly, sustained vowels with higher inherent emotion e. g., *Smiling* and Sad, were more influential and related to the level of the BAI, seen both from the acoustic analysis, e. g., those with high anxiety having lower STD $F_0$ for *Smiling* than those with low anxiety and vice versa for Sad. This is particularly interesting, as it suggests that speech can capture states of poor wellbeing better than neutral or positive levels and that emotionality is a strong indicator of this. Further, this may relate to the known ability for acoustic features to model states of highly arousal emotion better in general, in most scenarios, the sustained *Smiling* vowel was more easily modelled in individuals reporting higher levels of anxiety. In general, from the literature, the smile is known to alter the vocal tract more than other facial expressions and is said to be "heard as well as seen" [203]. Additionally, those with high anxiety often find emotion regulation a challenge and exaggerate their emotional expression [94], possibly leading to more substantial speech variance. Given these relations to facial expression, it would be valid to compare these results to other modalities, particularly video-derived features.

Another aspect that was being evaluated was the nature of the target itself, being derived from a self-report (**RQ-3**). As there are no perceived or objective marker-based labels available to compare to in this case, only assumptions can be made. However, it would seem that as an index, the BAI was indicative of the state of anxiety, given the consistencies in the findings across both experiments. One aspect which was less useful in the context and would better from a more profound analysis was the specific Has vs No symptoms groupings, and it would appear that in general, these results were less telling about the difference in the groups. This could be due to the nature of the questions being that they are requesting information about the subjects experienced in the last month, and analysis of the subjects during a target's period of anxiety where these symptoms are prominent at the time may show to be fruitful.

## 4.3 Continuous States of Emotional Wellbeing

In this section, two experiments are conducted, focusing primarily on continuous recognition of dimensional states of emotion from speech during lower-levels of wellbeing. The first experiment is a *classical* continuous speech emotion recognition paradigm, where the subjects targeted are within a public speaking scenario. The second experiment focuses on predicting a physiological adapted emotional target from speech and multimodal fusion during TSST. Proposing a novel and potentially more optimal representations of continuous emotional arousal in the context of stress and recognition of states of emotional wellbeing. These experiments are based largely on two published works, namely [25], and [24], and as with the previous these experiments address the following RQs in the following manner:

- **RQ-1:** Validating the use of audio-based features for targeting states of emotional wellbeing. In these experiments this is explored more specifically in relation to continuous dimensional models for emotion, and within varied scenarios which may present emotion differently.
- **RQ-2:** Exploring the use of audio in a uni- vs multimodal setting. Validating the overall benefit of audio as it pertains to continuous emotion (arousal and valence) and emotional wellbeing.
- **RQ-3:** Data scarcity in the context of computational analysis for emotional wellbeing is another aspect which is explored in both experiments. The first relates more specifically to the data itself, applying an audio-specific data augmentation method, and the second focuses on the annotation of continuous emotion, and presents an alternative in the case lower agreement between raters, or the need for additional raters.

### 4.3.1 Emotion During Public Speaking

In modern society, public dissemination is a useful tool for knowledge-sharing. However, having a fear of public speaking means that some individuals avoid this opportunity. Public speaking can provoke disorders, including Generalised Anxiety Disorder (GAD), and acute stress [204], both having a substantial effect on short-term wellbeing [205]. Furthermore, cultural differences in regards to an individual's response to the fear of public speaking have been researched, with markers including varied heart and speech rates [204]. To this end, observing emotional states during public speaking allows for a strong indication of the overall state of wellbeing [206], particularly as research has shown that an individual's typical emotion production can change during public speaking [207]. With this in mind, biological

signals are not readily observable and require rather invasive methods to be continuously captured. Audio, however, can be observed non-invasively, and has shown to be a reliable indicator for an individual's state or trait [208, 61].

For the current study, the main goal is to evaluate if speech-based audio features are useful for recognition of emotion during a public speaking scenario. To explore this deep learning-based approach is applied, utilising a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) architecture with self-attention, to predict continuous dimensional emotion (arousal and valence). An attention mechanism is applied as this has shown to improve results for sequence-based tasks, including emotion recognition [209, 210], and may allow for more context to be captured during training of the LSTM-RNN. For these experiments the BioS-DB as described in Chapter 3 is utilised, and a series of models are trained on various acoustic features. Moussu et al. [211] have shown that speaking in front of others in ones non-native language may cause more fear. Motivated by this, as the BioS-DB includes individuals speaking in both German and English during a public speaking scenario, the data is also grouped with consideration to language and native or not German.

#### 4.3.1.1 Data and Procedure

As mentioned the BioSpeech Database (BioS-DB) is used for these experiments. The audio data is first converted to 16 kHz, 16-bit WAV for use with popular feature extraction toolkits. As the dataset is reasonably small, an explicit validation set is excluded in this case, and the data is partitioned in to speaker-independent train and test partitions (see Table 4.17).

For these experiments as with others, a feature-based machine learning approach is applied. The dataset contained a number of physiological signals, and so features were extracted from the audio with this in mind, applying a 1-second window size and a small hop-size of 62.5 ms (16 Hz) for all audio features. As with previous experiments in this chapter, both hand-crafted speech features, and spectrogram-based data representations are extracted from the speech signals (see Table 4.18).

As the dataset being used is reasonably small, and to infer the effect of overall data quantity in this context, data augmentation is applied to the spectrogram-based features. For this spectrogram representations are extracted from each audio file at a window-size of 1 second and a hop-size of 62.5 ms. The A Simple Data Augmentation Method for Automatic Speech Recognition (SpecAugment) method [118] which masks portions of frequency and time from each extracted spectrogram is then applied. These spectrograms are then fed as an input the pre-trained CNN based architecture provided by the DeepSpectrum toolkit and features vectors are extracted from the augmented spectrogram images directly. With this approach the training set is then increased by a factor of 2.

For the target itself, a gold-standard for the emotion (arousal and valence) labels was calculated between the three raters utilising the EWE method. As described in Chapter 3, EWE and has been applied repeatedly on emotion-based datasets [60]. When fusing the signals, the mean inter-rater agreement across all speakers in the BioS-DB from the three annotators was 0.47 and 0.36 (based on a range of [0,1]) for arousal and valence, respectively. For the experiments, the gold-standard emotion labels were re-scaled to [-1,1] based on the maximum possible value.

Table 4.17: The speaker-independent partitions created for the BioS-DB, reporting (#) of native German (GER) and Non-native German (NonGER) speakers.

| # | **Train** | **Test** | $\sum$ |
|---|---|---|---|
| GER | 25 | 5 | 30 |
| NonGER | 7 | 5 | 12 |
| $\sum$ | 32 | 10 | 42 |

Table 4.18: An overview of the extracted features, used within these experiments. For further detail on each of the sets, see Chapter 3.

| **Feature Set** | **Modality** | **Dimensions** |
|---|---|---|
| eGeMAPS | Audio | 88 |
| DeepSpectrum | Audio | 4 096 |

#### 4.3.1.2 Experimental Settings

As the target for these experiments is continuously rated arousal and valence, and the audio signal is sequential in nature, an LSTM-RNN based is utilised as a regressor. The network consists of one recurrent LSTM layer with 128 units, a self-attention sequence layer, with a sequence-wide window, and sigmoid activation, and the output is then fed into a feed-forward layer which provides the predictions. For each sub-set of data the model is trained for 5 epochs with a batch size of 64 using the Adam optimiser and a learning rate of 0.001.

In an endeavour to explore the ability for capturing context, and also due to the relatively small hop-size applied for feature extraction, the input data in reshaped to sequences of 20 feature vectors (1.25 seconds). An alternative training strategy is applied for these experiments, where the model is updated in an iterative manner per speaker. To avoid potential speaker bias caused by this training method, several models are trained, and for each the order of speakers is shuffled.

Two types of language-based models (German and English) are also trained using the two acoustic feature sets (eGeMAPS and DeepSpectrum). For model testing, the speakers are grouped into Native-Germans speaking German (GER-GER), Native-Germans speaking English (GER-ENG), Non-Germans speaking German (NonGER-GER), and Non-Germans speaking English (NonGER-ENG), as well as from all test speakers together (All). To evaluate the prediction accuracy, the Concordance Correlation Coefficient (CCC) is utilised as the evaluation metric, given that it is established in the field of SER [212].

#### 4.3.1.3 Discussion of Results

Table 4.19: The test set results (reporting CCC) for continuous recognition of (A)rousal and (V)alence from the BioS-DB. Results obtained from the mean of all test speakers in that language-based grouping, across the 5 best-performing models trained on both English (ENG) and German (GER) languages. Emphasised results for arousal indicating a CCC $\geq$0.3, and for valence CCC $\geq$0.1. Results with * are discussed.

| | | **GER-GER** | | **GER-ENG** | | **NonGER-GER** | | **NonGER-ENG** | | **All** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CCC | **Train** | A | V | A | V | A | V | A | V | A | V |
| eGeMAPS | ENG | .072 | .045 | .165 | **.102*** | **.403** | **.279*** | **.582*** | **.175*** | .269 | **.130*** |
| | GER | .075 | .069 | .147 | .072 | **.382** | **.148** | **.370** | .088 | .203 | .084 |
| DeepSpectrum | ENG | .046 | .074 | .117 | .045 | .233 | .089 | **.349** | .063 | .147 | .037 |
| | GER | -.003 | .064 | -.004 | .021 | **.471** | .078 | **.339** | .010 | .114 | .028 |
| DeepSpectrum + | ENG | .172 | .056 | .018 | **.393** | .283 | **.308** | .156 | **.244** | .158 | **.226** |
| SpecAugment | GER | -.050 | **.192** | .096 | **.419** | **.640*** | **.491** | **.387** | **.296** | **.344** | **.334** |

An overview of the results for all experimental paradigms is given in Table 4.19. Where significant differences are discussed, this is based on the predictions from all speakers and a mean of all models and proceeds an evaluation of normality using a Shapiro-Wilktest [198]; a two-tailed T-test is also calculated, and the null hypothesis is rejected at a level of $p < 0.05$.

The results in Table 4.19 show that language appears to play a notable role for emotion recognition in this context. The best Native-German correlation for arousal is .260 CCC for the model trained on English, and tested on Germans speaking English (GER-ENG). As well as this, The German only models (GER-GER) have consistently negligible correlations as compared to NonGER-GER. Furthermore, looking at the Non-German results, a promising increase in CCC is observed, particularly for the English trained models. Indeed, in this paradigm, the best valence result comes from NonGER-GER, with .279 CCC. This result suggests that the positive to negative dimension of emotion (which is typically a challenge for audio modelling) is captured more easily when individuals are speaking in their non-native language, possibly due to higher levels of anxiety [213]. A result which is slightly agreed

upon with the GER-ENG valence score of .102 CCC, and even from the NonGER-ENG valence results of .175 CCC – further analysis of native English is a challenge, as BioS-DB there are only 2 native English speakers.

As discussed in the earlier experiment's focused on anxiety (see Section 4.2), it may be that the more neutral expression, caused by the neutral transcript of the text 'The North Wind and the Sun', which is limiting the ability to model aspects of valence. In other words, more vibrant expression may be captured better by the feature representations, however this is not commonly expressed when speaking publicly.

For the prediction of arousal, the best correlation is .582 CCC, and comes from the eGeMAPS English model, when testing on NonGER-ENG. Across most of the testing paradigms, the hand-crafted eGeMAPS features perform better than DeepSpectrum for this task. However, through the use of data augmentation, more stable results across all results in general are seen, with the best result for arousal of .640 CCC obtained in the NonGER-GER grouping. This suggests that data augmentation is suitable here, to improve robustness of results, and further establishes the findings in regard to the language groupings, as similar patterns of behaviour between the feature sets is seen. In this regard, for the GER-GER model trained on English there is an improvement with data augmentation, which would also point to the language dependency of the model and the task itself.

In summary, the results from these experiments show that audio is suitable for modelling arousal, with some promising results for valence in the context of public speaking, and that the data augmentation method proposed was particularly useful in obtaining a more robust result. For further research, it would be of interest to perform feature selection with the eGeMAPS features, to explore which features from this set perform highly in this context. As well as this, various audio-based augmentation approaches, such as additive noise, may also improve the robustness of these results. The window-based training approach also appears to have been effective for this use case, but additional hyperparameter optimisation would be extremely fruitful, particularly to explore the overall benefit that larger windows of context may have on model performance.

### 4.3.2 Physiologically-Adapted Emotion During Stress

Physiological and emotional responses can coincide during a stressful situation [214], and the degree of correlation has shown to be dependent on factors including underlying psychological traits and states, e. g., social desirability, or physiological dispositions, e. g., brain morphology [215]. During a stress-inducing situation, heart-rate, and breath become varied [216], along with the voice [112] (which is related strongly to perceived affect [217]). To this end, signals such Electrodermal Activity (EDA) (as known as SC) – described as a

psycho-physiological indication of emotional arousal [218] – correlate with an individual current perceived emotional state, specifically during high states of arousal [219].

Within the field of affective computing, recognition approaches to predict continuous states of emotion frequently utilise the two-dimensional Circumplex Model of Affect [96], observing the arousal (activation) and valence (positivity) of perceived emotion. However, as emotion is a subjective state of being, multiple raters must continuously annotate, which is time-consuming and costly. Further to this, the method to obtain a robust agreed-upon signal from multiple raters (gold standard) remains an ongoing research question, with several methods available (see Chapter 3, Section 3.1.3).

With this in mind, research into the fusion of physiological signals for use with perceived emotional signals is limited. Although physiological signals are utilised as features [220], there has been minimal research on a combined physiological and perceived arousal gold standard. Recently, in the 2021 edition of the Multimodal Sentiment Analysis in Real-life Media Challenge (MuSe), the signal of arousal was fused with EDA and used as a prediction target for the MuSe-Physio sub-challenge [15]. The baseline result from this was 0.3 CCC stronger than the arousal only *MuSe-Stress* sub-challenge when performing a late-fusion of audio and video-based features. Furthermore, the text-based features (typically less helpful for recognition of arousal) also improved when targeting EDA fused with arousal.

To explore this idea further, as the scenario is a stress induced situation, in this set of experiments the Ulm-TSST data as described earlier and applied in earlier experiments will be used to explore the fusion of EDA,BPM, and Rate of Respiration (RESP) with arousal ratings. As well as a number of the aforementioned benefits, e. g., reducing the amount of raters needed, or allowing for the addition of more raters, given this pseudo-professional setting, it can be considered that the utilisation of physiological signals (a more objective marker for arousal) may be of more use here, as perceived arousal may be suppressed to make a better impression towards the interviewer [218]. Furthermore, research has shown that the '*illusion of transparency*' can mean that alterations in speech are more prominent to the speaker than the audience [221], suggesting that continuous physiological signals may be more valuable for observation generally than perceived emotion. For these experiments, several multimodal features are extracted from audio, video, and textual transcriptions and a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) architecture is used as a regression algorithm, following a similar training procedure as outlined earlier in the chapter (see Section 4.1.1.1).

Table 4.20: An overview of the speaker-independent partitions, including number (#) of speakers and total duration of the data splits across Train, (Devel)opment and Test partitions for the sub-set of the Ulm-TSST dataset.

| | **Train** | **Devel** | **Test** | $\sum$ |
|---|---|---|---|---|
| # | 33 | 9 | 11 | 53 |
| **hh: mm: ss** | 2:45:29 | 0:45:32 | 0:55:33 | 4:26:36 |

#### 4.3.2.1 Data and Procedure

The full Ulm-TSST dataset consists of recordings from 110 German-speaking individuals (ca. 10 hours), which are annotated for the continuous dimensions of emotion (valence and arousal). For the experiments, a sub-set of the dataset is used including 53 speaker, reduced due to various artefacts during data processing. The data is in a speaker-independent train, development, and test partitioning, with balanced speaker demographics across the partitions (see Table 4.20). Only EDA, BPM and RESP are used for the physiological signals, and each is down-sampled to 2 Hz (to match the arousal ratings) and smoothed, applying a Savitzky–Golay filter, to reduce irrelevant, fine-grained artefacts in the signal.

To fuse the rating with the physiological signal a continuous annotator fusion technique Rater-Aligned Annotation Weighting (RAAW), first presented in [15] is used (see Chapter 3, Section 3.1.3). This method essentially first aligns the signals applying GCTW and then fuses them with consideration to the weighting of each rater using the EWE method.

In Table 4.21 the mean and standard deviation for the inter rater agreement when calculating each gold standard is given. For the experiments, six gold standards are created. Annotator 1 ($A_1$) and 2 ($A_2$), were selected as the consistent signals, as the correlation was strong between these two raters (see Figure 4.8). Where there are two annotators ($A_1$, $A_2$) and a physiological signal, the benefit of removing an annotator who is suboptimal (in other words, the annotator with the lowest agreement) is explored. Where two arousal raters are used with all physiological signals, this is exploring the advantage of using physiological signals to bring the rating in the gold standard up to five, see Figure 4.9 to observe the signal behaviours before and after fusion.

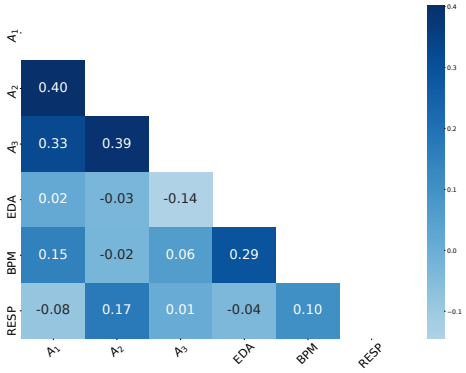



Figure 4.8: Correlation matrix between each individual signal across all speakers in the training set of the Ulm-TSST dataset. A stronger correlation is seen between (A)nnotator 1 and 2 with a number of the physiological signals also showing correlation between other signals.

Table 4.21: The mean ($\mu$) and standard deviation ($\sigma$) for inter-rater agreement between (A)nnotators 1-3 and the combinations of physiological signals, as Pearson correlation coefficient ($r$). Prior to EWE after GCTW.

| $r$ | $\mu$ | $\sigma$ |
|---|---|---|
| $A_1$,$A_2$,$A_3$ | .173 | .191 |
| $A_1$, $A_2$, EDA | **.230** | .241 |
| $A_1$,$A_2$ + BPM | .158 | .187 |
| $A_1$,$A_2$ + RESP | .108 | .134 |
| $A_1$,$A_2$,$A_3$, EDA,BPM | .119 | .155 |
| $A_1$,$A_2$,$A_3$, EDA,RESP | .088 | .120 |
| $A_1$,$A_2$,$A_3$, BPM,RESP | .070 | .097 |
| $A_2$, $A_2$, EDA, BPM, RESP | .127 | .123 |
| EDA, BPM, RESP | .197 | .149 |

The approach for these features is feature-based, and a combination of acoustic, vision, and textual-based features are applied. As speech is strongly linked to perceived arousal, two feature sets are used in this case, and these are used based on the better performing sets from the original Ulm-TSST baseline, described in [15]. This time not using the hand-crafted features (see Table 4.22), with further details including the applied alignment given in [15], as well as the earlier description in Section 4.1.

#### 4.3.2.2 Experimental Settings

Given the time-dependent nature of this task, an LSTM-RNN based architecture is utilised. Extensive hyperparameter optimisation is applied, including, as with previous experiments, the hidden state dimensionality of $32, 64, 128$ and the numbers of layers $1, 2, 4$, as well as learning rate $0.0001, 0.001, 0.005$. As with the experiments of Section 4.1.1.1, for the training processes, the features and labels of every input are further segmented via a windowing approach [184], and in this case window-size of 300 steps (150 seconds) and a hop-size of 50 steps (25 seconds) is again found to be optimal.

To observe the benefits of multimodal approaches, a decision-level (late) fusion is applied to evaluate the co-dependencies of the modalities. The experiments are restricted to the best performing features from each modality only. For decision-level fusion, separate models are trained individually for each modality.

Table 4.22: An overview of the extracted features used within these experiments. For a description of the FAU, BERT, and VGGFace features, see [15], and for further detail on audio-based features see Chapter 3.

| **Feature Set** | **Modality** | **Dimensions** |
|---|---|---|
| DeepSpectrum | Audio | 4 096 |
| VGGish | Audio | 128 |
| BERT | Textual | 768 |
| VGGFace | Vision | 512 |

#### 4.3.2.3 Discussion of Results

To explore the benefit of physiological-based arousal and perceived arousal fusion, the extensive results for the computational prediction experiments conducted are given in separate tables for ease of discussion (see Table 4.23 and Table 4.24). As an evaluation metric for these experiments, CCC is employed, as is typical for emotion recognition tasks, and to better compare to the initial baseline results obtained for the Ulm-TSST dataset [15].

For the results in Table 4.23, it can be seen that the perceived arousal only ($A_1$-$A_3$) score is strong, particularly from a multimodal approach where at best .506 CCC is achieved on the test set, from late-fusion of audio and video-based features. However, looking at the uni-modal approaches for $A_1$-$A_3$, as expected given the pseudo-professional scenario of the TSST, audio-only features capture the perceived arousal to a lesser degree compared to VGGFace. Furthermore, as is typical for arousal prediction tasks, the uni-modal textual features perform worst, obtaining .2118 CCC on the test set.

This finding in relation to the pseudo-professional scenario is similar to what is disused in the previous public speaking experiments (see Section 4.3.1), regarding the emotionality of the speech from experiments in Section 4.2 being more easily modelled. In the last experiments where language appeared to play a decisive role, as native speakers were less easily modelled when speaking in front of others, this is a similar scenario here (for arousal only gold standard), and suggests a need for adaption of the arousal gold standard to be more representative of the speakers' current state.

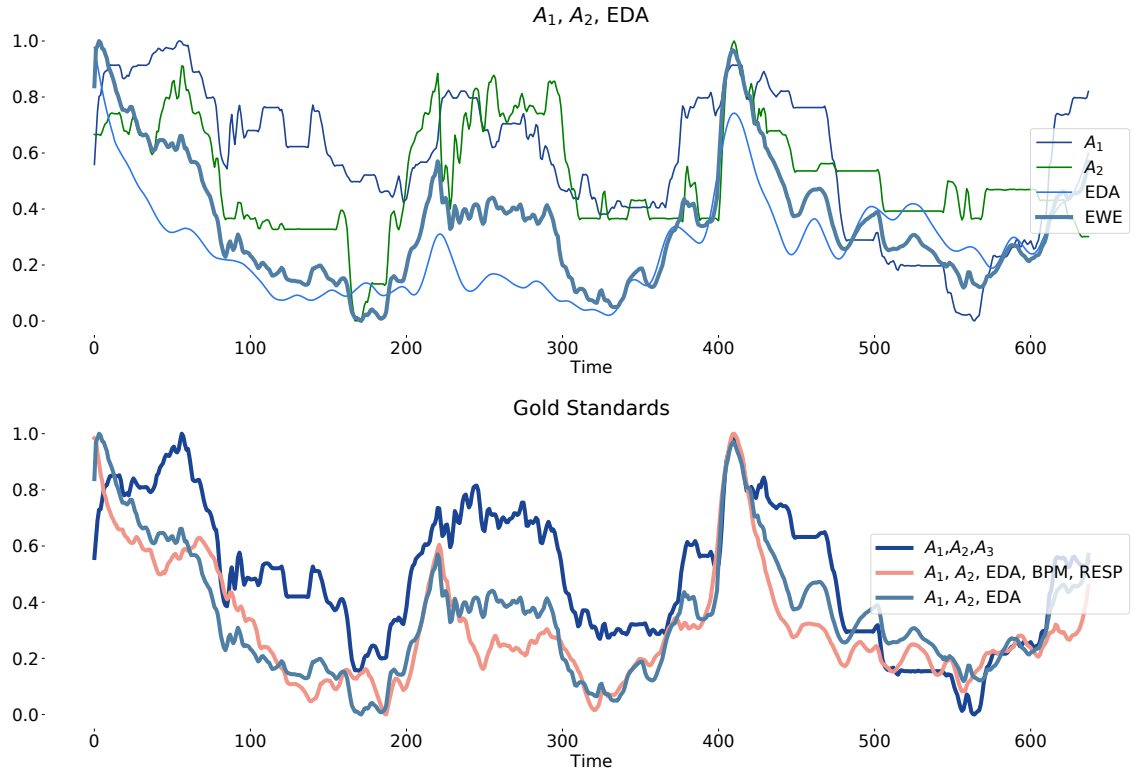



Figure 4.9: An example of the gold standard creation for subject #9 from Ulm-TSST. Above, (A)nnotator $A_1$, $A_2$ + EDA ($\sigma = .217$), and below, a comparison of three gold standards.

Table 4.23: The CCC results for prediction of an arousal only and single physiological signal adapted arousal gold standard, on the (devel)opment and test partitions. Utilising (V)ision: VGGFace, (A)udio: DeepSpectrum, VGGish (VGGish), and (T)ext: Bidirectional Encoder Representations from Transformers (BERT). Reporting the best result from hyperparameter optimisation, as well as reporting the mean ($\mu$) across all feature sets for a given signal. Best test scores are emphasised.

| **Perceived** | $A_1$,$A_2$,$A_3$ | | $A_1$,$A_2$ | | $A_1$,$A_2$ | | $A_1$,$A_2$ | |
|---|---|---|---|---|---|---|---|---|
| **Physiological** | | | EDA | | BPM | | RESP | |
| CCC | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| VGGFace | .3025 | .3813 | .3216 | .3959 | .4805 | .3771 | .1869 | **.3745** |
| DeepSpectrum | .2826 | .3060 | .3366 | .4031 | .1649 | .2327 | .0382 | .0977 |
| VGGish | .2127 | .2856 | .3493 | .4210 | .3156 | .3313 | -.0079 | .1716 |
| BERT | .1341 | .2118 | .2431 | .2402 | .0567 | .1037 | .1063 | .1802 |
| A + V | .4638 | **.5062** | .4506 | **.5103** | .4640 | .3889 | .3196 | .3108 |
| A + T | .3240 | .3841 | .3821 | .3470 | .3044 | .3205 | .1396 | .2032 |
| V + T | .2526 | .4668 | .3442 | .4213 | .4735 | **.4202** | .3443 | .2871 |
| A + V + T | .3476 | .4965 | .4186 | .4987 | .4458 | .4104 | .3811 | .3036 |
| **$\mu$ of All** | – | .3798 | – | .4047 | – | .3231 | – | .2411 |

Table 4.24: The CCC results for prediction of physiological signal adapted arousal gold standard with up to five signals, on the (devel)opment and test partitions. Utilising (V)ision: VGGFace, (A)udio: DeepSpectrum, VGGish, and (T)ext: BERT. Reporting the best result from hyperparameter optimisation, as well as reporting the mean ($\mu$) across all feature sets for a given signal.

| **Perceived** | $A_1$,$A_2$,$A_3$ | | $A_1$,$A_2$,$A_3$ | | $A_1$,$A_2$,$A_3$ | | $A_1$,$A_2$ | |
|---|---|---|---|---|---|---|---|---|
| **Physiological** | EDA,BPM | | EDA,RESP | | BPM,RESP | | EDA,BPM,RESP | |
| CCC | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** | **Devel** | **Test** |
| VGGFace | .3694 | .4062 | .3995 | .3941 | .3637 | .4306 | .4704 | .4707 |
| DeepSpectrum | .3089 | .3861 | .1841 | .3807 | .2527 | .2046 | .3683 | .3832 |
| VGGish | .4851 | .5164 | .0901 | .3985 | .2689 | .3649 | .5161 | .4712 |
| BERT | .1999 | .0542 | .2733 | .2393 | .1210 | .0922 | .3568 | .3344 |
| A + V | .5666 | **.6157** | .3630 | **.3947** | .4722 | **.4432** | .6674 | .5025 |
| A + T | .5089 | .3677 | .3249 | .1777 | .3295 | .2817 | .5570 | .4357 |
| V + T | .4839 | .3783 | .3836 | .2301 | .3738 | .3881 | .5916 | **.5355** |
| A + V + T | .5895 | .4596 | .4028 | .3470 | .4086 | .4230 | .6669 | .5055 |
| **$\mu$ of All** | – | .3980 | – | .3203 | – | .3285 | – | .4548 |

Table 4.25: The CCC results for prediction of physiological signal only gold standard, on the (devel)opment and test partitions. Utilising (V)ision: VGGFace, (A)udio: DeepSpectrum, VGGish, and (T)ext: BERT. Reporting the best result from hyperparameter optimisation, as well as reporting the mean ($\mu$) across all feature sets for a given signal.

| **Physiological** | EDA,BPM,RESP | |
|---|---|---|
| CCC | **Devel** | **Test** |
| VGGFace | .5679 | **.5838** |
| DeepSpectrum | .4189 | .5157 |
| VGGish | .3197 | .4613 |
| BERT | .2909 | .3842 |
| A + V | .5030 | .5728 |
| A + T | .4175 | .5586 |
| V + T | .4386 | .5594 |
| A + V + T | .4623 | .5639 |
| **$\mu$ of All** | – | **.5250** |

Table 4.26: The mean ($\mu$), standard deviation ($\sigma$) of reported CCC for a selection of test results given in Table 4.23, Table 4.24 and Table 4.25, which include EDA, BPM, or RESP.

| | $A_1$,$A_2$,$A_3$ | **inc. EDA** | | **inc. BPM** | | **inc. RESP** | |
|---|---|---|---|---|---|---|---|
| CCC | | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| VGGFace | .3813 | .4167 | .0364 | .4212 | .0396 | .4175 | .0424 |
| DeepSpectrum | .3060 | .3883 | .0101 | .3017 | .0965 | .2666 | .1402 |
| VGGish | .2856 | **.4518** | .0527 | .4210 | .0872 | .3516 | .1279 |
| BERT | .2118 | .2170 | .1174 | .1461 | .1273 | .2115 | .1018 |
| **Late-Fusion** | | | | | | | |
| A + V | .5062 | **.5058** | .0903 | .4876 | .0972 | .4128 | .0810 |
| A + T | .3841 | .3320 | .1096 | .3514 | .0663 | .2746 | .1162 |
| V + T | .4668 | .3913 | .1263 | .4305 | .0722 | .3602 | .1339 |
| A + V + T | .4965 | .4527 | .0733 | .4496 | .0427 | .3948 | .0888 |

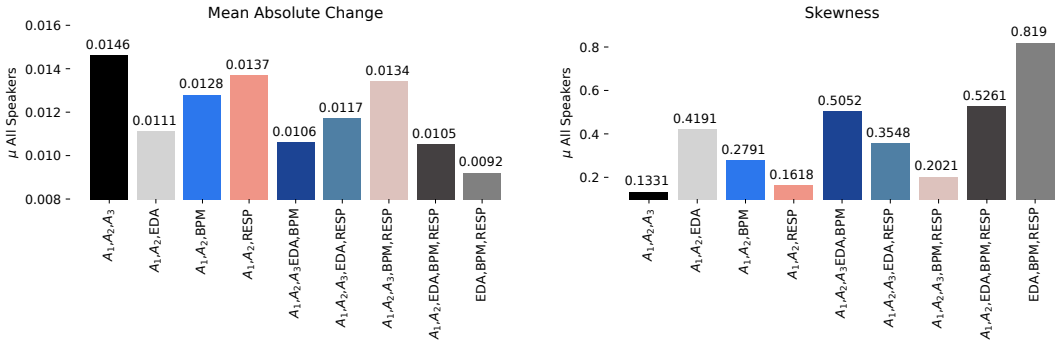



Figure 4.10: The Mean Absolute Change (MAC) and skewness for the mean ($\mu$) of all speakers, from each gold standard signal utilised within experiments.

With this in mind, continuing to look at Table 4.23, in general a slight improvement across features when incorporating a physiological signal is seen. Of interest, a ca. .3 CCC improvement for BERT features when utilising the EDA signal. Typically, perceived arousal is challenging for textual-based features, as seen from the perceived arousal baseline. However, at best for BERT features when predicting the combined $A_1$, $A_2$, EDA, BPM, RESP signal, a CCC of .334 is obtained, which is .1 above the $A_1$-$A_3$ baseline. When observing the mean across experiments, including EDA Table 4.26, it is confirmed that EDA is the strongest physiological signal for the BERT features.

For the audio features, it is also shown that a more robust result is obtained over VGGFace when utilising EDA, suggesting that the behaviour of EDA is present in the voice, making this gold standard more attainable for the speech-based features. Similar behaviour for BPM and RESP fusion is also obtained; however, this is not as consistent as EDA results, as shown from the more consistent mean results in Table 4.26. Furthermore, there are lower results for the $A_1$, $A_2$ BPM, and $A_1$, $A_2$ RESP results, compared to the $A_1$-$A_3$ baseline.

Furthermore when discussing audio features, all gold standard approaches, which include EDA, report an improvement, and up to .471 CCC is obtained by VGGish features, where all physiological signals are utilised. In Figure 4.9, it can be seen that the two physiological-adapted gold standards follow a similar trend to the arousal baseline gold standard. However, there is a slightly reduced standard deviation for this example, with .24 for $A_1$-$A_3$ compared to .22 for $A_1$, $A_2$, EDA, and .203 for the $A_1$, $A_2$, EDA, BPM, RESP signal. This may suggest that better results are obtained from a smoothing effect when the perceived arousal is fused with physiological signals.

To further analyse this smoothing effect, it can be seen in Table 4.25 that the results are consistently higher than the arousal only baseline when utilising the physiological only

(EDA-RESP) signal. With a standard deviation of .157 for the same example in Figure 4.9, this does lean more toward being a factor in the improvement of the results.

However, the MAC, and skewness are also extracted from each of the gold standard signals across all speakers in Figure 4.10. Although further investigation should be done here, there is an inherent difference in the MAC from $A_1$-$A_3$, and EDA-RESP, which is mirrored by the skew of the signals' distribution. Of promise, and perhaps opposing the smoothing effect, none of the physiological signal only results obtains higher than the best result when fusing with perceived arousal, i. e., .6157 CCC from $A_1$, $A_2$, EDA, BPM with audio and video feature fusion. This leads to the consideration that further investigation on this topic may be fruitful – particularly, as there is no reduction in results from physiological-adapted arousal fusion.

### 4.3.3 Conclusions

From the results of both these experiments, it is again clear that speech derived features can be applied to target states emotional wellbeing (**RQ-1**). From the first experiments, this was shown prominently, particularly by the evaluation of native against non-native speakers, where arousal was modelled in a strong way, with valence being modelled to a moderate degree via the implementation of spectrogram based data augmentation.

In the case of the gold standard adaption task, findings have shown that in most cases, the EDA signal can improve recognition of arousal, specifically interesting for textual based features, but also aiding acoustic features, for which the less (perceived) aroused speech behaviours may have been a challenge. Furthermore, this finding was in general found to be more beneficial when replacing a poor rater with a substantially low inter-rater agreement (**RQ-3**). There was less of an improvement from BPM or RESP signals alone, however, when fused with EDA, various feature sets did see improvements, with the best score obtained from a fusion of perceived arousal with BPM and EDA of up to .6157 CCC based on late-fusion of audio and video features.

In general, these works show again that audio can work as a uni-modal signal for modelling various states of emotional wellbeing (**RQ-2**), although there is again certainly an improvement in this context when fused with vision-based features. In this same way from the first experiments the spectrogram data augmentation strategy also appears to be robust for these types of application.

## 4.4 Audio Generation for Speech Emotion Recognition

As data is sparse in this area of research, generative networks can be applied as an augmentation approach to generate novel samples of speech data. In this section, first the efficacy of this in relation to emotional speech is explored, and proceeding to this a method for evaluating the generated samples is then presented. These experiments are based largely on two published works by the author, firstly [170], where WaveNet was applied for the first time in the context of generating emotional speech. The limitations of this work [170] were then addressed in a later publication [26], which is the main focus for the current experiments. These experiments address the following RQs in the following manner:

- **RQ-3a:** To explore the validity of generating emotional speech with the well-established WaveGAN architecture, and the efficacy of applying generated data as an augmentation strategy to combat the issues pertaining to data scarcity in the realm of emotional wellbeing.
- **RQ-3b:** Given that data augmentation is a popular method for tackling data scarcity, an evaluation method is proposed which allows for a more interpretable evaluation of generated samples, for which there is currently limited methods. Exploring the ability of this method to discuss attributes including, similarity, diversity, and plausibility.

### 4.4.1 Data and Procedure

In both experiments, a subset of the GEMEP dataset will be used (see Chapter 3 for details). When processing the raw audio, first it is converted to 16 kHz, 16 bit, mono, WAV format, and split into three (speaker-independent) folds (see Table 4.27). The partitioning chosen is applied for all experiments and considers a balance between classes and speaker demographics as best possible.

A WaveGAN model is trained to generate the new audio data, as first proposed in [172] (see Chapter 3 for details). The WaveGAN model is trained using the data partitioned into the first fold (F-1). In the GEMEP dataset, samples are of varied length. As WaveGAN requires fixed-length data, 1-second chunks selected from the samples during training.

The WaveGAN applied was trained using the default parameters described in [172] for 100 000 training steps. For these experiments, samples are generated until the quantity is equal to the classes within the source training data (total of 526 1-second samples). From a qualitative evaluation of the generated audio, the samples do have similar attributes as

the source speakers[2]. Of note, as is typical for GAN generated audio, there is a noise artefact in the high-frequency range, which are also visible in the extracted spectrograms (see Figure 4.11). In future work, the inclusion of a processing step (denoising, or low-pass filtering) to remove such artefacts may be of value for comparison.

To compare any results obtained with WaveGAN generated data, several low-resource audio augmentation approaches are also applied, namely, time-shifting and additive noise, and spectrogram warping with time and frequency masking (e. g., the SpecAugment method [118]). These types of augmentation are chosen for the audio to give a broad range of representations to compare. As can be seen in Figure 4.11, the time-shift representation is most similar to the source; and subjectively, the additive noise or SpecAugment approaches are the most dissimilar. The total number of samples from the training set of the GEMEP dataset is duplicated for each of these augmentation approaches. The audio signal is moved by a maximum of 0.5 seconds from the end of the signal for time-shifting the audio samples, selecting the value for time-shift randomly for each sample. For additive noise, white noise is injected at a SNR of 1 dB, from the amplitude from the source. The SpecAugment augmentation approach is applied to the spectrograms directly, for more detailed information on this approach, see Chapter 3 and Section 3.1.2.1, as well as [118].

Table 4.27: The speaker-independent folds used for both experiments, reporting quantity for each of the four emotional classes utilised from a sub-set of the GEMEP dataset.

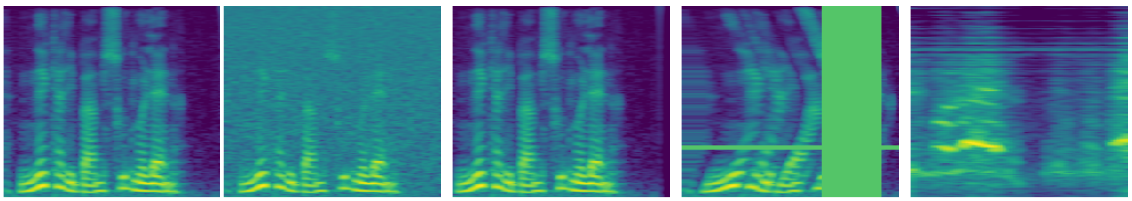
| | **Fold-1** | **Fold-2** | **Fold-3** | $\sum$ |
|---|---|---|---|---|
| Speakers (M:F) | 6 (3:3) | 2 (1:1) | 2 (1:1) | 10 |
| Pleasure | 60 | 18 | 12 | 90 |
| Anger | 60 | 18 | 12 | 90 |
| Elation | 48 | 18 | 24 | 90 |
| Sadness | 48 | 18 | 24 | 90 |
| $\sum$ | 216 | 72 | 72 | 360 |

#### 4.4.1.1 Experimental Settings

For both the experiments, an adaptation of the prototypical network first presented in [167] is proposed[3]. The model is first used directly as a classifier, and then the embedding space which the network has learned is explored more deeply. For further detail on the Prototypical

[2]To listen to a selection of the generated samples for each of the four classes visit https://shorturl.at/mwDZ1 accessed 09.2021.

[3]github.com/EIHW/prototypical-network-audio-evaluation accessed on: 09.2021



(a) Original (b) Noise (c) Time-shift (d) SpecAugment (e) WaveGAN

Figure 4.11: Mel-spectrogram representation (with a range of 0–8 kHz) for source and augmented audio samples, take from a sample in the *anger* class.

network see Chapter 3, Section 3.3.1.6, as well as [167], particularly for further details of specific terminology not explicitly defined herein.

For the embedding function, which learns data representations based on the four classes, a CNN based architecture is applied. The model consists of four convolutional blocks, each with a convolutional layer, batch normalisation, a ReLU activation function, and a max-pooling layer. The first three blocks have a $3 \times 3$ filters, 64 channel output, and the last $3 \times 3$ blocks have a reduced channel output of 32. For the first two convolutional blocks, a $2 \times 2$ max-pooling layer is applied, and for the third layer, the max-pooling is increased to $4 \times 4$. As a measure to avoid overfitting the network, 40 % drop out before the last convolutional block is applied. The model is trained with the Adam optimiser applying an initial learning rate of $10^{-3}$, which is halved every epoch of 100 episodes. An episode can be referred to as a mini-batch. For all experiments, training is stopped after ten epochs. The episodic sampling approach mitigates the class imbalance in the data, as samples are evaluated in randomly selected class pairings. Given this, the F-score ($F_1$) is reported as an evaluation metric. Each model is evaluated five times, with the highest $F_1$ is reported.

As an input source for the classifier, spectrogram images are extracted (see Figure 4.11) from the raw audio. The spectrograms are extracted with a pixel dimensionality of $256 \times 256$ (a dimension within a common range as applied to other speech emotion studies [222]), and the colour map *viridis*. A maximum frequency of 8 kHz is applied to the spectrograms to reduce the presence of high-frequency non-speech related activity.

### 4.4.2 Emotional Speech Generation

There are many advantages to generating new audio data computationally, mainly the scarcity of actual data, particularly in the speech emotion domain [170]. The time-dependent nature of audio makes sourcing and annotating such data an extremely time-consuming process [223], and so generative models such as GANs [172] or DARNs [174] can be used as an augmenta-

tion method. However, only limited research has been done on the ability of a generative network to learn more salient aspects of audio, such as emotion in speech. Data augmentation is one quantitative approach for evaluating the plausibility of generated audio [224], and therefore, for these first experiments, the data generated is applied to the training set (Fold-1) of the source data, and compared with the three other data augmentation approaches.

#### 4.4.2.1 Discussion of Results

Table 4.28: The results obtained for the data augmentation experiments. Training a prototypical network with source data augmented with Additive (Noise), SpecAugment, Time-shifting, and WaveGAN data. Fold-3 of the source data is used for the test evaluation. Reporting $F_1$ as an evaluation metric.

| **Fold-1** | **Test** $F_1$ |
|---|---|
| Source Baseline | 60.4 |
| Source + Noise | 61.8 |
| Source + SpecAugment | 61.0 |
| Source + Time-shift | 60.2 |
| Source + WaveGAN | **63.9** |

When observing the results in Table 4.28, although improvement with data augmentation is minimal, the WaveGAN data does appear to be able to generate emotionality as there is a slight improvement when applying this data as an augmentation approach compared to all others, at best results of 63.9 % $F_1$. This is particularly the case when compared to other approaches, where the worst performing was time-shift, which reports results slightly lower than the baseline. These lower results for time-shift may be explained by a high degree of similarity in the embedding space and their limited diversity. In general, these results do establish the plausibility of applying WaveGAN to the task of generating emotional speech, however there is limited interpretation that can be made from these results alone.

### 4.4.3 Evaluating Generated Audio

In the previous experiments, it would seem that emotionality can be generated. However, the interpretability from a quantitative perspective of how the samples behave within the embedding space is limited when applying the generated data only as a data augmentation method. Therefore in the proceeding experiments, an evaluation framework based on a prototypical

network is outlined. The approach evaluates similarity and diversity, which, when applied to domains including emotional speech, may allow for a more human-interpretable discussion and more fine-grained evaluation.

The core aspect of this is approach is to harness the latent space learnt by the prototypical network, which includes prototype representations of each class from the generated and source audio. As previously, other augmentation approaches are used as an anchor to compare the learnt prototypes too.

#### 4.4.3.1 Additional Experimental Setting

When augmenting a training set, considering the similarity and diversity of the new data to the already known data is a necessary factor [225]. To this end, to explore the use of prototypical networks as an evaluation method for these aspects of generated data, two core experiments are performed, which are described as follows:

*Generated data similarity:* As a first-step to observe the similarity of the generated samples, two prototypical networks are trained on the source data, one which uses Fold-1, and the other using a concatenation of Folds 2 and 3 (see Table 4.27). These models are then evaluated with data from Fold-1, for each of the data augmentation types. Samples are classified based on the euclidean distance between support class prototypes and query samples, and therefore it can be assumed that samples with higher distance (lower similarity) to the support class prototypes will be miss-classified.

*Pairwise-embedding space diversity:* To investigate the diversity of the generated data, the distances between samples in the trained model's embedding space are analysed. A representation of a sample in the embedding space is a data point. It is assumed that two similar samples lead to similar representations and, therefore, a small distance of points in the embedding space. Counter to this, two samples from a diverse dataset are expected to lead to separation in the embedding space. To explore this, point-pairs are built to match each generated data point with its closest source data point according to a calculation of the Euclidean distance. Finally, the mean Euclidean distance between all points in a pair is calculated. As a reference point, the generated data is compared to source data from a different fold as well as and the source data from the same fold.

#### 4.4.3.2 Discussion of Results

The results for similarity and diversity experiments, are given in, Table 4.29 for *generated data similarity* and Figure 4.12 for *pairwise-embedding space diversity*. For ease of discussion, the

Table 4.29: Reporting the $F_1$ (%) obtained for the generated data similarity experiment. Training models on (F)old-1, and Fold-2+3 source data, and evaluating with all data combinations – Source, Additive (Noise), SpecAugment, Time-shift, and WaveGAN.

| **Trained on** | **Evaluated with** | **Test score** | **Trained on** | **Evaluated with** | **Test score** |
|---|---|---|---|---|---|
| Source-F1 | Source-F1 | 95.6 | Source-F2+F3 | Source-F1 | 59.3 |
| | Noise-F1 | 61.4 | | Noise-F1 | 46.0 |
| | SpecAugment | 77.9 | | SpecAugment | 48.3 |
| | Time-shift-F1 | **87.8** | | Time-shift-F1 | **57.5** |
| | WaveGAN-F1 | 53.1 | | WaveGAN-F1 | 43.6 |

results will be outlined individually. As a baseline understanding, the prototypical networks trained for evaluating prototype similarity are reporting accuracy's above chance level (25 %) (see Table 4.29). These results suggest that the network can differentiate between the four classes to a reasonably high degree. For example, in Figure 4.13, it can be seen that the class-prototypes from the Source-F2+3 experiments as a t-Distributed Stochastic Neighbour Embedding (t-SNE) representation appear to have definition within the embedding space, with source data clusters very close to the class prototype.

Specifically for evaluating the augmentation types, the time-shifting approach appears to have a consistently strong test accuracy, 87.8 and 57.5 $F_1$, for F1 train and F2+F3 train, respectively. This finding would confirm that this is the most similar to the source data. For all other augmentation types between the models, the results are less clear, and when evaluating with unseen data (Fold-2+3 model), Noise, SpecAugment and WaveGAN fall within a similar range. Furthermore, the SpecAugment approach appears to perform reasonably well, which may be due to retaining considerable aspects of the source, which is then harnessed by the convolution layers. For the WaveGAN samples, the model can classify the data above chance level, which does show that some emotionality must have been learnt. However, results are lower than that of all augmentation approaches. This low performance still shows promise for the WaveGAN samples, as it shows that it is in the range of the source class prototypes but perhaps has higher diversity in the embedding space.

The next experiments focused on evaluating diversity, and for these experiments, the embedding space is analysed from the models of the previous experiment (generated data similarity). The mean Euclidean distance between an augmented query set data points and the closest source query set data point in the prototypical (Source-F1, and Source-F2+3) embedding space is calculated. As a reference, the same measure for the source support

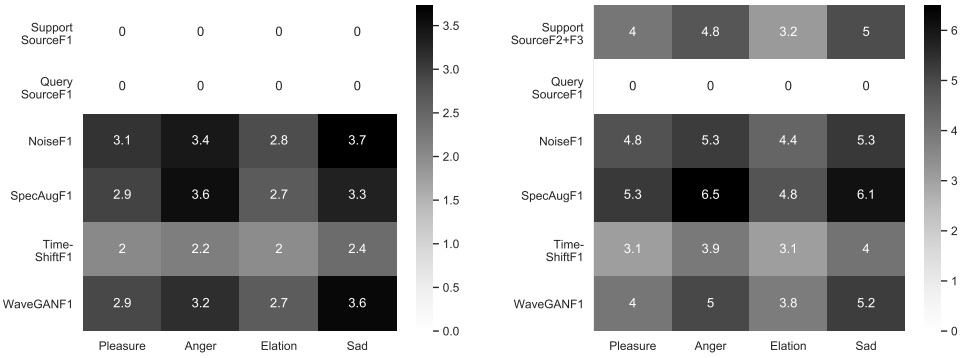



Figure 4.12: A heat map representation of the results for the pairwise-embedding space diversity experiments. Left is the Source-F1 trained model, and right is the Source-F2+3 trained model. Reporting the mean of each augmentation type's absolute Euclidean distance from the source query samples.

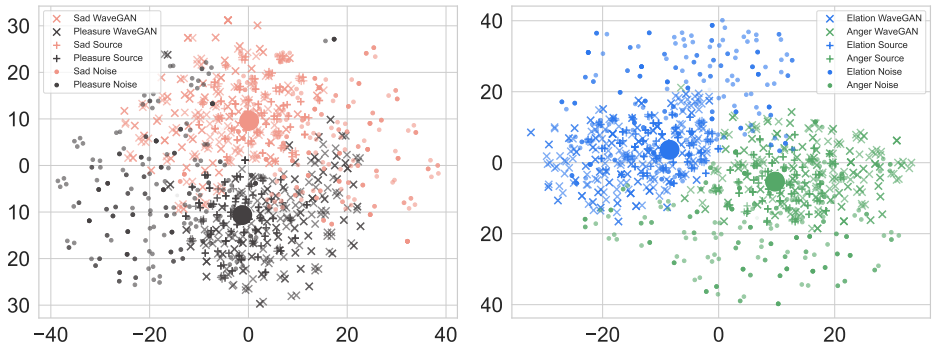



Figure 4.13: t-SNE representation (learning rate= 20 and perplexity= 90) for the classes of sad vs pleasure (left), elation vs anger (right) and classes of interest from the WaveGAN generated, Source Fold-1, and Noise augmented data within the prototypical embedding space of the F2+F3 model, where the prototype is based on the source data support set.

samples and the source query samples is also provided. The prototypical network was trained with Fold-1 and Fold-2+3, respectively.

Any augmentation technique is based on the source data of Fold-1, therefore, implies an inherent dependence within the data. Figure 4.12 shows that the time-shift augmentation has the smallest average pair-distance, which is in line with the assumption that this augmentation technique only slightly modifies the source data. Therefore, the time-shift adds the least diversity to the data. The right plot of Figure 4.12 implies that the WaveGAN samples show a similar average pair-distance as the independent source support samples taken from F2+3.

It would appear in general that the WaveGAN samples add a similar level of diversity to the source query data as additional independent source support data might. Finally, the noise and SpecAugment generated samples show a higher average pair-distance than the

other approaches, especially in the case of SpecAugment samples, which suggests that these approaches add more diversity than others. However, as it seems unlikely that any suggested augmentation method adds more variance than independent source data, which would improve augmentation results, the higher average pair-distance might also result from a distortion of the source data. The left side of Figure 4.12 depicts the pairwise distance where source query and source support data are identical. This again shows a very similar diversity trend for augmented samples.

### 4.4.4 Conclusions

From these experiments, it becomes clear that generative models are able to produce emotionality to a degree which is distinguishable by a classifier can distinguish (**RQ-3a**). This finding is promising as it supports the use of generative networks to tackle the issue of data scarcity in speech-based monitoring of emotional wellbeing. The initial results where plausibility was evaluated, and the generated data is used to augment the source training set, in theory, presented this. However, with these experiments minimal interpretation was available.

In this regard, the prototypical network-based framework was applied to evaluate two additional aspects of the generated samples within an embedding space – similarity and diversity. Despite the complexity of spectrogram images and the limited training data used, both of those aspects were able to be discussed in the context of the emotional speech samples generated by the WaveGAN. The similarity and diversity results, in fact, support that the initial plausibility experiments i. e., data augmentation were obtained for the WaveGAN from the data being neither too similar nor too diversely spread in the embedding space.

More rigorous testing is needed, including human listening studies to compare the behaviour of the learned embedding space to that of a humans perception. Furthermore, comparing these results to other quantitative evaluation metrics such as the inception score may offer further insight. As a continued theme, the nature of training the WaveGAN model was somewhat brute-force, and the variance in speaker characteristics within the training set may have an implicit effect on the generated samples – exploring personalised model training may be advantageous. Of most promise from the current experiments, it appears that an understanding of all evaluation criteria can be obtained (**RQ-3b**), and through various visualisations of the embedding space, this approach can allow for a more interpretable representation of the generated audio.

# Concluding Remarks

## 5.1 Summary

This thesis covers an extensive range of topics in the area of computer audition for emotional wellbeing. As a result, several conclusions can be discussed in regards to the initial research questions introduced in in Chapter 1 Section 1.1. The main focus was to explore how computer audition-based methods can be applied to monitor and understand states of emotional wellbeing (**RQ-1**). This was aimed primarily at speech-based computational paralinguistic analysis, and, with emotional wellbeing being the main target of interest, a use case of improving empathic AI was introduced.

Overall, it is clear from the experiments that the audio signal can be applied to target an extensive range of states of emotional wellbeing (**RQ-1**). These include physiological markers of stress, specific emotional states such as anxiety, and specific emotion-inducing scenarios e. g., public speaking. The audio signal is particularly suited to monitoring highly aroused states, including stress and anxiety, in an array of scenarios. Occasionally though, the literature suggests that individuals may not express states such as stress outwardly, in certain situations [218]. With this in mind, a physiologically adapted representation for emotional arousal was explored, and this showed that audio was much more able to target a gold-standard signal which included the objective, continuous measure of electrodermal activity along with arousal. This was due, in part, to this signal being more indicative of an individual's current state than a perceived rating of arousal. Similarly, for many of the experiments, a common theme was the variance in speaker behaviour and the overall need for more personalised approaches. Personalisation would be particularly meaningful for speech-driven approaches given the inherent variance in expression and the core anatomical differences that are unavoidably present during a vocalisation. This would certainly be a necessary factor in the context of improving empathic AI, where generalisation becomes slightly more of an ethical concern.

Furthermore, it has been shown in the experiments that, in general, audio consistently performs to a high degree as a uni-modal signal (**RQ-2**). The research in this thesis shows that many emotional-driven states of wellbeing can be robustly targeted with the audio signal.

Nevertheless, when speech is missing, particularly in higher states of emotional distress, audio inherently becomes less valuable. As such, the experiments performed supported that audio can benefit from a fusion of vision and textual-based features. However, audio, in a sense, has the ethical upper hand over video for instance, in that it can be captured in a pseudo-anonymous way, due to the inferiority of auditory vs visual memory [226]. Furthermore, as humans are generally visually dominant, video can be more challenging in regards to subject privacy, despite it being extremely valuable for a number of tasks. So, in this sense, the audio signal is a valuable resource as not only is it able to model a variety of states of wellbeing, it can also be integrated in a ubiquitous and non-invasive way.

Leading on from this, during the development of this thesis, several datasets were collected and provided to the academic community (see Table 3.1) (**RQ-3**). However, in general, the majority of experiments performed are based on reasonably small-scale datasets, which is a major limitation for computer audition. To explore this, the experiments of Chapter 4 Section 4.1.1.1 utilised three datasets from differing acoustic environments and found that the targets in question could be modelled to a similar degree across the varied acoustic characteristic, e. g., room size, microphone usage. This was promising as it shows that, particularly in the area of computational emotional wellbeing, through collaboration with expert external research groups, several datasets can be combined for exploration. As a side note, this was less successful in the context of physiological signals, as we saw that the characteristics of the signal varied across datasets in such a way that made modelling targets e. g., BPM in a multi-domain manner challenging. This was likely caused by the varied equipment and site-specific calibration. Similarly, as it pertains to data scarcity (**RQ-3**), in the experiments of Chapter 4 Section 4.4, it was shown that for the purpose of data augmentation, generative networks are suited to the generation of classes of emotional speech, improving classification accuracy when incorporated with the training set. One limitation of this was the ability to interpret generated audio samples efficiently. As such, an approach in which the embedding space could be discussed was presented, and it was found that similarity, diversity, and plausibility could all be discussed within that process.

Many of the experiments conducted, and the fundamental tools available for this field, have required interdisciplinary collaboration (**RQ-4**). This ranges from the production of hand-crafted feature sets that remain a strong representation of the state of emotional wellbeing to the experimental design and deriving a ground truth. Nevertheless, through this research, the primary input from an interdisciplinary perspective was on sourcing psychologically backed data; this is a required aspect for human-derived data and was vitally informative for all outcomes discussed herein. However, when it comes to the modelling itself, within the community, this remains more of a 'black-box' process to external researchers. A

large component of the research made herein was written to be as accessible as possible. As it pertains to the empathy of AI, it is clear that moving forward, accessible knowledge sharing for psychologists, health care professionals, and other related fields, needs to continue to improve the specific vocabulary being utilised.

## 5.2 Ethical Considerations

Several ethical considerations relate to the above summary and these should be outlined as they relate to this thesis. This is particularly important given the human focus of this thesis and the potential for unethical commercial exploitation that this type of research may be exploited for. As it pertains to CP in [227] the authors outline a road-map for good ethical practices in the field of CP but also with consideration to the broader field of AI. When interfacing with this type of research, there are a number of ethical considerations, including storage, anonymisation, and privacy (e. g., consent for data usage), which should always take precedence. However, it has become clear throughout the development of this doctoral research that there are three core ethical issues that the community should continue to consider, 1. bias-free and representative data 2. interpretable decisions 3. interdisciplinary collaboration. These will be discussed in more detail in the following section.

### 5.2.1 Bias-Free and Representative Data

One of the biggest challenges for the fields connected to machine learning is the representative nature of the data that is being modelled. This is particularly relevant when the target is human; however, the implications of poorly representative data are not limited to explicitly human-derived data [228, 229]. Over the years, there have been several prominent media articles that have indicated a great deal of bias occurring in a number of domains, e. g., racial bias in health care [230].

There are several forms of bias that are often discussed in relation to AI, including historical bias, interaction bias, latent bias, and *selection bias*. *Selection bias* is particularly related to AI data as this occurs when the data pooled for analysis is not representative of the larger population, e. g., skewed towards a specific gender, age, or employment status. This would lead to misrepresentation rather than generalisation. A typical example of this type of bias is in regards to gender [231]. In [232] the authors found, for example, that models tend to have a bias towards a particular gender even when a dataset is balanced. This could indicate lower level architecture-based biases, derived from early decision-making by the developer [233]. In the case of audio, this could be caused by less dominant feature-based

behaviours from certain demographics. *Selection bias* is particularly essential to address when referring to models developed for human interaction; as such, it is highly relevant to the discussion of empathy in AI. From data-based decision making, a bias can propagate through a system's architecture, leading to poor accuracy on a diverse population. Lack of true generalisation is particularly problematic for domains such as health, where, in critical cases, this may result in a breach of patient safety [234].

A core contributing factor to bias in AI is the management of data. Data is often sourced in a centralised less flexible way, where individuals present a unified data source to a central server i. e., missingness or alternative inputs outside of a controlled scale are often not possible. This approach creates an arguably homogeneous representation of the target population where only the static aspects of a given individual are considered and certain participants may not be represented [235]. This is consistent with the concerns that most AI models are based on Western, Educated, Industrialised, Rich, and Democratic (WEIRD) societies [227]. Research needs to expand this to underrepresented regions which may not have the resources to manage the potential impact of this, particularly where models are being integrated into ubiquitous smart devices, which are not available global[1].

### 5.2.2 Interpretable Decisions

From both a technical and applied perspective, machine learning, when applied to human behaviour modelling, needs to be interpretable by the general public, particularly those interfacing with it. Within deep learning, networks are seemingly becoming more complex, and it is crucial in the AI evolution to continue to develop strategies for understanding the internal decision making of machine learning algorithms [236]. Similarly, those responsible for developing such technology need to be transparent about the reasons for developing these models and what exactly is behind them.

With the fast-paced environment of technological advancements, it is not difficult for researchers in machine learning to glance over the finer details of the reasons behind their conclusions, or why their model performed so well on a particular dataset. From the perspective of speech, performing statistical analysis of the acoustics LLDs and applying feature importance type strategies would all contribute to an improved, more easily discussed, early-stage understanding. However, many approaches in the audio domain are now working on raw audio, which can make understanding the learning process somewhat less accessible. So, although the data may represent one way of understanding why certain behaviours were modelled better than others, researchers in the field of eXplainable Artificial Intelligence

[1] 80%$\geq$ of the global population now own a smart device. Bank my Cell 2021, accessed: 09.2021

(XAI) should continue to explore other strategies for observing the decisions of the networks themselves. For example, in recent years, considerable work has focused on attention-based methods for interpretability. Given the nature of attention layers, they are deemed to be interpretable as they allow for context to be considered (based on the weighting of a given input in a sequence), and observation of activations within the embedding space throughout the training period [165].

Even with the technical approaches being developed, it remains vital that general advancements in AI and machine learning are communicated to the general public in a manageable and consumable way. When working with domains such as emotion understanding, this is particularly vital as currently, the knowledge gap is growing, and many individuals report having fear for such methodologies. This is primarily due to a lack of genuine communication by 'Big Tech' in regards to why a particular AI application is integrated or what exactly it means when it is. One option could be similar to the type of tax relief certain governments afford to companies who contribute to charities, this could be similar for AI dissemination. However, implementing a financial incentive is perhaps not the best practice as this may force a kind of cognitive empathy, where researchers work on interpretability as a means to an end, rather than understanding its true impact. Instead, a change in the overall ethos of the AI community is needed, and it is a moral right for users to be afforded the agency to understand the types of decisions that are behind the technology they interface with daily.

### 5.2.3 Interdisciplinary Collaboration

In a similar way to the above, it is critical that those within the technical domain work alongside those with greater expertise and understanding of the human condition. Returning to the idea of Artificial General Intelligence (AGI), interdisciplinarity has shown to be a necessary step forward for this next phase of AI [89]. Interdisciplinarity is particularly valuable as the literature suggests that infrastructures developed in this way more easily tackle ethical concerns relating to acceptance, bias, and trust [89].

Social acceptance of AI integration is necessary for its success and long-term adoption by the public. A range of aspects, including cultural and environmental impact, need to be considered, and various experts should provide knowledge on the target areas. For example, the synthesised voice of bus announcements not representing the community to which it speaks may have a negative impact on those communities. A closer analysis of the voice that best represents that community would be more ethically considerate, and more likely to be accepted [237]. In this way, facilitating interdisciplinary collaboration between engineers and linguists or sociologists would aid more considerate and empathic AI.

Similarly, knowledge of bias often requires contributions from experts with non-technical backgrounds, and an approach for facilitating discussion between fields of research would be a valuable next step. For example, within the machine learning community, techniques such as *few-shot learning* have received more attention in recent years [238] due to the advantages that they pose for computational efficiency. However, perceptual-based biases pose difficulties for such approaches [239], and discussion from experts of the targeted domains may help understand the bias at an earlier stage.

In a similar way, research focused on improving empathy of AI is another area that will strongly benefit from non-technical input. With this in mind, understanding the communication strategies between differing fields speaking different "languages" (i. e., anthropology and engineering), is an important area to focus on in itself, as this will lead to improved trust from the public and more empathic interactions with AI systems. In this way, due to historical stereotypes, AI continues to have lower levels of trust by the general user. For example, these are users who, without an understanding of the vocabulary of the field, may not be able to grasp the core concept of such machine learning networks. Through a better collaboration with various academic researchers, communicating AI to the general public may also improve which, in turn, will help to build trust. Of note, trust was shown to improve when interactions appeared to have more empathy, particularly for voice assistance's [88].

## 5.3 Limitations

Although the findings and contributions from this doctoral research are substantial, there are a number of limitations which should be discussed. These are highlighted in an endeavour to be transparent and offer building blocks for other research to build on.

*The variance in the manifestation for states of emotional wellbeing across speakers* within the datasets applied in the experiments of this thesis has not been considered in depth. Although the overarching goal for machine learning is to generalise a given dataset, it should be considered that in the context of wellbeing, emotion and mental health can manifest in extremely differently in each person, across modalities, and depending on the situation. For example, as indicated, aroused emotions can be suppressed by certain individuals more than others, leading to differing perceptions in the degree of arousal that the individual is facing. However, it is more than this; in the context of wellbeing, low-level human attributes, such as gender, age, or even further prosodic characteristics, would also be beneficial to condition on. Clustering features to understand larger groupings of subjects within a dataset may be one area to focus on [27].

*Speaker enhancement and general denoising.* Improving the quality of the raw audio signal is a large area of speech-based machine learning research that is not focused on within this thesis. This area would, from a data processing perspective, be extremely valid to explore regarding benefits when targeting states of emotional wellbeing. However, developing machine learning models that can learn representation in noisy and unclear data does also have its merits. For data cleaning and enhancing, one aspect is therapist patient, cross-talk, and more generally, quiet speech, which may be masked by louder acoustic activity. Furthermore, with the rise in remote counselling and life-coaching, improving signal quality for general interpretability for the health care practitioner would be meaningful, in the same way as it would be for a listening machine. Similarly, when discussing the generative approaches for emotional speech, additional denoising may further improve the quality of the generated samples, which often retain several digital artefacts.

*There is limited use of End-to End modelling* within this thesis, although this is not completely absent. It should be noted that end-to-end models are popularly applied in general and, on large datasets, the results are impressive. However, in this case, large really does mean very large. As mentioned for computer audition-based datasets, particularly as it pertains to emotional wellbeing, this scale of data is much more difficult to come by. Furthermore, as it pertains to the use of deep learning in general, it has been noted in [240] that the overall opaque machine learning models should be avoided when interpretability and analysis of the phenomena is critical. This has, of course, been more of a focus for this current work.

*The creation of a gold standard in the context of emotional wellbeing* remains a challenge in itself for the affective computing community, with perceived vs self-assessed emotions inherently representing the data in different ways. Within the community, and for particular use cases, research is still needed to determine the model for emotion that is most useful for understanding a true human expression. In other words, should models be attempting to replicate (potentially biased) human perception? Or, should they be attempting to understand expressions which are more of a challenge for humans to perceive, yet relate more to how an individual might truly feel? This is particularly relevant for more atypical human expression and, as such, it is extremely important to develop models that are not biased towards typical emotional expression.

## 5.4 Outlook

From the research conducted, there are a number of avenues that should be targeted by the community when continuing to develop computer audition-based applications for emotional wellbeing. As well as the current literature and prominent trends, the outlook from the perspective of this thesis is also based on some of the previously discussed limitations.

Given the aforementioned variance in speaker characteristics and the manifestation of emotion-based states of individual wellbeing, as well as the potential and consideration for bias in models, one technical area which would be extremely beneficial for this domain is the work being developed in personalised machine learning and speaker-adaptation. Work has been developing in this area for a number of years, and a fundamental approach is feature adaptation i. e., normalisation of features, on a per-subject basis [241], where the purpose is to reduce the overall speaker variability, whilst preserving the discrimination between emotional classes. Furthermore, a number of more unsupervised strategies have been presented recently which may also be beneficial to explore, including the application of transfer learning from a pre-trained model, which is then adapted to single speakers [242].

Similarly, much of the research being developed in machine learning-based computer audition topics originate from computer vision where the datasets are much larger, and the input is potentially easier to manage. As such, research needs to focus on more audio-specific approaches which consider the nuance of audio. Similarly, as the results here support, audio in certain circumstances does benefit from a fusion with other modalities, and novel methods are being developed which explore a more meaningful multimodal representation in the context of speech emotion recognition. This is particularly prominent when it comes to multi-level or multi-head attention mechanisms which appear to be extremely valuable [243].

Finally, given the extremely sensitive nature of emotional wellbeing, in the context of empathic AI, greater attention needs to be given by the AI community in regards to the representation of the emotion targets themselves. In other words, higher dimensionality, and consideration to the vast differences in emotional expression, need to continue to be explored. Prominent work in this direction includes a deep understanding of emotional representation in vocal burst [244]. However, the literature still appears to require further interdisciplinary collaboration and longer, more time-consuming studies to evaluate longer in duration samples, particularly as pertains to dyadic pairings, and interpersonal relationships in general. Such approaches may provide further insights into understanding the effects of perceived vs felt emotion in the context of wellbeing.

# Acronyms

**A | B | C | D | E | F | G | H | I | K | L | M | N | O | P | R | S | T | U | V | X**

**A**

**AGI** Artificial General Intelligence.

**AI** Artificial Intelligence.

**ALC** Alcohol Language Corpus.

**AMDF** Average Magnitude Difference Function.

**ANN** Artificial Neural Network.

**auDeep** Unsupervised Learning of Representations from Audio with Deep RNNs.

**AudioSet** An Ontology and Human-labeled Dataset for Audio Events.

**B**

**B-LSTM** Bidirectional Long Short-Term Memory.

**BAI** Beck Anxiety Inventory.

**BERT** Bidirectional Encoder Representations from Transformers.

**BioS-DB** BioSpeech Database.

**BPM** Beats per Minute.

**BPTT** Back-propagation Through Time.

**BVP** Blood Volume Pulse.

**C**

**CCA** Canonical Correlation Analysis.

**CCC** Concordance Correlation Coefficient.

**CE** Cross-Entropy.

**CLIA** Chemiluminescence Immunoassay.

**CMU-MOSEI** CMU Multimodal Opinion Sentiment and Emotion Intensity.

**CNN** Convolutional Neural Network.

**ComParE** Computational Paralinguistics Challenge.

**CP** Computational Paralinguistics.

**CSLE** Cognitive Load with Speech and EGG Corpus.

**CTW** Canonical Time Warping.

**D**

**DAC** Düsseldorf Anxiety Corpus.

**DARN** Deep Auto-regressive Networks.

**DC-GAN** Deep Convolutional Generative Adversarial Networks.

**DeepSpectrum** Spectrogram-based Feature Extraction from Audio Data with Pre-trained Convolutional Neural Networks.

**DELFIA** Dissociation-Enhanced Lanthanide Fluorescence Immunoassay.

**DEMoS** Database of Elicited Mood in Speech.

**DNN** Deep Neural Network.

**DTW** Dynamic Time Warping.

**E**

**EDA** Electrodermal Activity.

**eGeMAPS** extended Geneva Minimalistic Acoustic Parameter Set.

**EMO-DB** Berlin Database of Emotional Speech.

**EmoNET** A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition.

**EWE** Evaluator Weighted Estimator.

**F**

**F-FNN** Feed-Forward Neural Network.

**FAU** Facial Action Units.

**FAU-TSST** Friedrich-Alexander-Universität-Trier Social Stress Test.

**FFT** Fast Fourier Transform.

**G**

**GAD** Generalised Anxiety Disorder.

**GAN** Generative Adversarial Network.

**GCTW** Generalised Canonical Time Warping.

**GEMEP** Geneva Multimodal Emotion Portrayals Corpus.

**GRU** Gated Recurrent unit.

**H**

**HNR** Harmonic-to-Noise Ratio.

**HPA** Hypothalamic Pituitary Adrenal axis.

**HR** Heart Rate.

**I**

**IEMOCAP** Interactive Emotional Dyadic Motion Capture.

**K**

**kNN** k-Nearest Neighbours.

**KSS** Karolinska Sleepiness Scale.

**L**

**LLDs** Low-Level Descriptors.

**LSTM** Long Short-Term Memory.

**LSTM-RNN** Long Short-Term Memory Recurrent Neural Network.

**M**

**MAC** Mean Absolute Change.

**MAE** Mean Absolute Error.

**MBC** Munich Bio-voice Corpus (MBC).

**MFCCs** Mel-Frequency Cepstral Coefficients.

**MSE** Mean Square Error.

**MSP-Podcast** Multimodal Signal Processing Podcast Dataset.

**MuSe** Multimodal Sentiment Analysis in Real-life Media Challenge.

**MuSe-CaR** The Multimodal Sentiment Analysis in Car Reviews Dataset.

**N**

**NLP** Natural Language Processing.

**nmol/L** Nanomoles per Litre.

**O**

**OCD** Obsessive-Compulsive Disorder.

**openSMILE** open-source Speech and Music Interpretation by Large-space Extraction.

**openXBOW** open-Source Crossmodal Bag-of-Words Toolkit.

**P**

**PTSD** Post-Traumatic Stress Disorder.

**R**

**RAAW** Rater-Aligned Annotation Weighting.

**RASTA** Relative Spectra.

**RAVDESS** The Ryerson Audio-Visual Database of Emotional Speech and Song.

**RECOLA** REmote COLlaborative and Affective interactions.

**Reg-TSST** Regensburg University-Trier Social Stress Test.

**RESP** Rate of Respiration.

**RMSE** Root Mean Square Error.

**RNN** Recurrent Neural Network.

**S**

**SC** Skin Conductance.

**SER** Speech Emotion Recognition.

**SEWA** Social Empowerment through Work Education and Action.

**SHS** Sub-Harmonic Summation.

**SinS-DB** Sincerity in Speech Database.

**SLC** Sleep Language Corpus.

**SLEEP** Düsseldorf Sleepy Language.

**SMOTE** Synthetic Minority Oversampling Technique.

**SNR** Signal-to-Noise Ratio.

**SpecAugment** A Simple Data Augmentation Method for Automatic Speech Recognition.

**SPL** Sound Pressure Level.

**STFT** Short-time Fourier transform.

**SVM** Support Vector Machine.

**SVR** Support Vector Regression.

**T**

**t-SNE** t-Distributed Stochastic Neighbour Embedding.

**TSST** Trier Social Stress Test.

**TTS** Text-to-Speech.

**U**

**UAR** Unweighted Average Recall.

**UCL-SBM** UCL Speech Breath Monitoring Corpus.

**Ulm-TSST** Ulm University-Trier Social Stress Test.

**URTI** Upper Respiratory Tract Infection Dataset.

**USoMS** The Ulm State-of-Mind in Speech Corpus.

**USOMS-e** The Ulm State-of-Mind in Speech (Elderly) Corpus.

**V**

**VAD** Voice Activity Detection.

**VAE** Variational Autoencoders.

**VGG16** VGG16.

**VGGFace** VGGFace.

**VGGish** VGGish.

**X**

**XAI** eXplainable Artificial Intelligence.

# Bibliography

[1] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[2] W. Wang, *Machine Audition: Principles, Algorithms and Systems*. Hershey, PA, USA: IGI Global, 2010.

[3] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: A survey," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 9411–9457, 2021.

[4] N. Saleem, M. I. Khattak, and E. Verdú, "On improvement of speech intelligibility and quality: A survey of unsupervised single channel speech enhancement algorithms," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 2, pp. 78–89, 2020.

[5] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, "A survey: Neural network-based deep learning for acoustic event detection," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3433–3453, 2019.

[6] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[7] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[8] B. Schuller, A. Baird, A. Gebhard, S. Amiriparian, G. Keren, M. Schmitt, and N. Cummins, "New avenues in audio intelligence: Towards holistic real-life audio understanding," *Trends in Hearing*, vol. to appear, 2021.

[9] B. W. 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," in *Proceedings of INTERSPEECH, 21th Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 2042–2046, ISCA, 2020.

[10] E. D'Arca, A. Hughes, N. M. Robertson, and J. Hopgood, "Video tracking through occlusions by fast audio source localisation," in *Proceedings of ICIP 2013, IEEE International Conference on Image Processing*, (Melbourne, Australia), pp. 2660–2664, IEEE, 2013.

[11] A. Maccagnan, S. Wren-Lewis, H. Brown, and T. Taylor, "Wellbeing and society: Towards quantification of the co-benefits of wellbeing," *Social Indicators Research*, vol. 141, no. 1, pp. 217–243, 2019.

[12] R. Dodge, A. P. Daly, J. Huyton, and L. D. Sanders, "The challenge of defining wellbeing," *International Journal of Wellbeing*, vol. 2, no. 3, pp. 222–235, 2012.

[13] G. E. Coverdale and A. F. Long, "Emotional wellbeing and mental health: An exploration into health promotion in young people and families," *Perspectives in Public Health*, vol. 135, no. 1, pp. 27–36, 2015.

[14] I. Tautkute, T. Trzcinski, and A. Bielski, "I know how you feel: Emotion recognition with facial landmarks," in *Proceedings of CVPR2018, IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (Salt Lake City, UT, USA), pp. 1878–1880, IEEE, 2018.

[15] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proceedings of MuSe'21, 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Chengdu, China), ACM, 2021.

[16] B. Schuller, S. Amiriparian, G. Keren, A. Baird, M. Schmitt, and N. Cummins, "The next generation of audio intelligence: A survey-based perspective on improving audio analysis," in *Proceedings of ISAAR 2019, 7th International Symposium on Auditory and Audiological Research*, vol. 7, (Nyborg, Denmark), pp. 101–112, 2019.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of NIPS 2017, Advances in Neural Information Processing Systems*, vol. 30, (Long Beach, CA, USA), pp. 5998–6008, 2017.

[18] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," 2016. arXiv:1606.05328.

[19] A. Baird, N. Cummins, S. Schnieder, and B. W. Schuller, "An evaluation of the effect of anxiety on speech – computational prediction of anxiety from sustained vowels," in *Proceedings of INTERSPEECH, 21st Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 4951–4955, ISCA, 2020.

[20] E. Emerson, N. Fortune, G. Llewellyn, and R. Stancliffe, "Loneliness, social support, social isolation and wellbeing among working age adults with and without disability: Cross-sectional study," *Disability and Health Journal*, vol. 14, no. 1, 2021. Art. no. 100965.

[21] M. Richardson, H.-A. Passmore, R. Lumber, R. Thomas, and A. Hunt, "Moments, not minutes: The nature-wellbeing relationship," *International Journal of Wellbeing*, vol. 11, no. 1, 2021.

[22] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, and D. Silove, "The global prevalence of common mental disorders: A systematic review and meta-analysis 1980–2013," *International Journal of Epidemiology*, vol. 43, no. 2, pp. 476–493, 2014.

[23] A. Baird, S. Amiriparian, M. Berschneider, M. Schmitt, and B. Schuller, "Predicting biological signals from speech: Introducing a novel multimodal dataset and results," in *Proceedings of MMSP 2019, 21st International Workshop on Multimedia Signal Processing*, (Kuala Lumpur, Malaysia), IEEE, 2019.

[24] A. Baird, L. Stappen, L. Christ, L. Schumann, E. Messner, and B. W. Schuller, "A physiologically-adapted gold standard for arousal during a stress induced scenario," in *Proceedings of MuSe'21, 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Chengdu, China), ACM, 2021.

[25] A. Baird, S. Amiriparian, M. Milling, and B. W. Schuller, "Emotion recognition in public speaking scenarios utilising an LSTM-RNN approach with attention," in *Proceedings of SLT 2021, IEEE Spoken Language Technology Workshop*, pp. 397–402, IEEE, 2021.

[26] A. Baird, S. Mertes, M. Milling, L. Stappen, T. Wiest, E. André, and B. W. Schuller, "A prototypical network approach for evaluating generated emotional speech," in *Proceedings of INTERSPEECH, 22nd Annual Conference of the International Speech Communication Association*, (Brno, Czechia), pp. 3161–3165, ISCA, 2021.

[27] L. Stappen, L. Schumann, B. Sertolli, A. **Baird**, B. Weigel, E. Cambria, and B. W. Schuller, "MuSe-Toolbox: The multimodal sentiment analysis continuous annotation

fusion and discrete class transformation toolbox," in *Proceedings of MuSe'21, 2nd International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Chengdu, China), ACM, 2021.

[28] A. Baird, E. Coutinho, J. Hirschberg, and B. W. Schuller, "Sincerity in acted speech: Presenting the sincere apology corpus and results," in *Proceedings of INTERSPEECH, 20th Annual Conference of the International Speech Communication Association*, (Graz, Austria), pp. 539–543, ISCA, 2019.

[29] B. Schuller, S. Steidl, P. Marschik, H. Baumeister, F. Dong, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & self-assessed affect, crying & heart beats," in *Proceedings of INTERSPEECH, 19th Annual Conference of the International Speech Communication Association*, (Hyderabad, India), pp. 122–126, ISCA, 2018.

[30] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. S. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," in *Proceedings of INTERSPEECH, 20th Annual Conference of the International Speech Communication Association*, (Graz, Austria), pp. 2378–2382, ISCA, 2019.

[31] A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, B. M. Kudielka, S. Sturmbauer, N. Rohleder, E.-M. Messner, and B. Schuller, "Evaluating speech-based recognition of emotional & physiological markers of stress," *Frontiers in Computer Science, Human-Media Interaction*, vol. to appear, 2021.

[32] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–16, 2021.

[33] S. Dubnov, "Computer audition: An introduction and research survey," in *Proceedings of MM '06, 14th ACM International Conference on Multimedia*, (Santa Barbara, CA, USA), p. 9, ACM, 2006.

[34] K. Qian, X. Li, H. Li, S. Li, W. Li, Z. Ning, S. Yu, L. Hou, G. Tang, J. Lu, F. Li, S. Duan, C. Du, Y. Cheng, Y. Wang, L. Gan, Y. Yamamoto, and B. W. Schuller,

"Computer audition for healthcare: Opportunities and challenges," *Frontiers in Digital Health*, vol. 2, 2020. Art. no. 5.

[35] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Organization of hierarchical perceptual sounds," in *Proceedings of IJCAI'95, 14th International Joint Conference on Artificial Intelligence*, vol. 1, (Montreal, Canada), pp. 158–164, 1995.

[36] B. W. Schuller, *Intelligent Audio Analysis*. Springer, Berlin, Heidelberg, 2013.

[37] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[38] S. E. McAdams and E. E. Bigand, eds., *Thinking in Sound: The Cognitive Psychology of Human Audition*. Oxford, UK: Oxford University Press, 1993.

[39] E. Vincent, T. Virtanen, and S. Gannot, eds., *Audio Source Separation and Speech Enhancement*. Chichester, UK: Wiley, 2018.

[40] S. Amiriparian, S. Julka, N. Cummins, and B. Schuller, "Deep convolutional recurrent neural network for rare acoustic event detection," in *Proceedings of DAGA 2018, 44. Jahrestagung für Akustik*, (Munich, Germany), pp. 1522–1525, DEGA, 2018.

[41] A. Hazan, R. Marxer, P. Brossier, H. Purwins, P. Herrera, and X. Serra, "What/when causal expectation modelling applied to audio signals," *Connection Science*, vol. 21, no. 2–3, pp. 119–143, 2009.

[42] Y. Zhao, X. Xia, and R. Togneri, "Applications of deep learning to audio generation," *IEEE Circuits and Systems Magazine*, vol. 19, no. 4, pp. 19–38, 2019.

[43] M. Vlaming and L. Feenstra, "Studies on the mechanics of the normal human middle ear," *Clinical Otolaryngology & Allied Sciences*, vol. 11, no. 5, pp. 353–363, 1986.

[44] A. F. Rawdon-Smith and G. C. Grindley, "An illusion in the perception of loudness," *British Journal of Psychology*, vol. 26, no. 2, pp. 191–195, 1935.

[45] E. Asutay and D. Västfjäll, "Perception of loudness is influenced by emotion," *PLOS ONE*, vol. 7, no. 6, 2012. Art. no. e38660.

[46] P. Sorokowski, D. Puts, J. Johnson, O. Żółkiewicz, A. Oleszkiewicz, A. Sorokowska, M. Kowal, B. Borkowska, and K. Pisanski, "Voice of authority: Professionals lower their vocal frequencies when giving expert advice," *Journal of Nonverbal Behavior*, vol. 43, no. 2, pp. 257–269, 2019.

[47] A. Baird and B. Schuller, "Considerations for a more ethical approach to data in AI: On data representation and infrastructure," *Frontiers in Big Data*, vol. 3, 2020. Art. no. 25.

[48] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, S. Hantke, and B. Schuller, "The perception of emotions in noisified nonsense speech," in *Proceedings of INTERSPEECH, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), pp. 3246–3250, ISCA, 2017.

[49] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. M. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates," in *Proceedings of INTERSPEECH, 22th Annual Conference of the International Speech Communication Association*, (Brno, Czechia), pp. 431–435, ISCA, 2021.

[50] K. Sandmann, A. am Zehnhoff-Dinnesen, C.-M. Schmidt, K. Rosslau, R. Lang-Roth, M. Burgmer, A. Knief, P. Matulat, M. Vauth, and D. Deuster, "Differences between self-assessment and external rating of voice with regard to sex characteristics, age, and attractiveness," *Journal of Voice*, vol. 28, no. 1, 2014. Art. no. 128–e11.

[51] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. 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, 2016.

[52] L. Chen, S. Gunduz, and M. T. Ozsu, "Mixed type audio classification with support vector machine," in *Proceedings of ICME 2006, IEEE International Conference on Multimedia and Expo*, (Toronto, Canada), pp. 781–784, IEEE, 2006.

[53] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proceedings of ICASSP 2017, IEEE International Conference on Acoustics, Speech and Signal Processing*, (New Orleans, LA, USA), pp. 131–135, IEEE, 2017.

[54] A. Shewalkar, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235–245, 2019.

[55] Z. Peng, J. Dang, M. Unoki, and M. Akagi, "Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech," *Neural Networks*, vol. 140, pp. 261–273, 2021.

[56] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *Proceedings of ICASSP 2021, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7118–7122, IEEE, 2021.

[57] S. Pascual, J. Serrà, and J. Pons, "Adversarial auto-encoding for packet loss concealment," 2021. arXiv:2107.03100.

[58] S. T. Roweis, "One microphone source separation," in *Proceedings of NIPS 2000, Advances in Neural Information Processing Systems*, vol. 13, (Denver, CO, USA), pp. 793–799, 2000.

[59] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[60] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, "MuSe 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *Proceedings of MuSe'20, 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Seattle, WA, USA), pp. 35–44, ACM, 2020.

[61] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proceedings of INTERSPEECH, 10th Annual Conference of the International Speech Communication Association*, (Brighton, UK), pp. 312–315, ISCA, 2009.

[62] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proceedings of INTERSPEECH, 11th Annual Conference of the International Speech Communication Association*, (Makuhari, Chiba, Japan), pp. 2794–2797, ISCA, 2010.

[63] C.-M. Tsai, S.-L. Chou, E. N. Gale, and W. D. McCall, "Human masticatory muscle activity and jaw position under experimental stress," *Journal of Oral Rehabilitation*, vol. 29, no. 1, pp. 44–51, 2002.

[64] D. O'Shaughnessy, *Speech Communications: Human And Machine*. Reading, MA, USA: Addison-Wesley, 1987.

[65] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, 2021. Art. no. 1163.

[66] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proceedings of ICASSP 2020, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 911–915, IEEE, 2020.

[67] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proceedings of INTERSPEECH, 12th Annual Conference of the International Speech Communication Association*, (Florence, Italy), pp. 3201–3204, ISCA, 2011.

[68] F. Schiel, C. Heinrich, and S. Barfüsser, "Alcohol language corpus: The first public corpus of alcoholized German speech," *Language Resources and Evaluation*, vol. 46, no. 3, pp. 503–521, 2012.

[69] T. Bänziger and K. R. Scherer, "Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus," in *A Blueprint for Affective Computing: A Sourcebook and Manual*, pp. 271–294, Oxford, UK: Oxford University Press, 2010.

[70] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association*, (Lyon, France), pp. 148–152, ISCA, 2013.

[71] B. Schuller, F. Friedmann, and F. Eyben, "The Munich BioVoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *Proceedings of LREC '14, 9th Language Resources and Evaluation Conference*, (Reykjavik, Iceland), pp. 1506–1510, 2014.

[72] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & physical load," in *Proceedings of INTERSPEECH, 15th Annual Conference of the International Speech Communication Association*, (Singapore, Singapore), pp. 427–431, ISCA, 2014.

[73] T. F. Yap, *Speech production under cognitive load: Effects and classification*. PhD thesis, Electrical Engineering & Telecommunications, Faculty of Engineering, University of New South Wales, Australia, 2012.

[74] J. Krajewski, S. Schnieder, and A. Batliner, "Description of the upper respiratory tract infection corpus (URTIC)," in *Proceedings of INTERSPEECH, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), ISCA, 2017.

[75] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, cold & snoring," in *Proceedings of INTERSPEECH, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), pp. 3442–3446, ISCA, 2017.

[76] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, sincerity & native language," in *Proceedings of INTERSPEECH, 17th Annual Conference of the International Speech Communication Association*, (San Francisco, CA, USA), pp. 2001–2005, ISCA, 2016.

[77] M. Kawahara, D. A. Sauter, and A. Tanaka, "Culture shapes emotion perception from faces and voices: Changes over development," *Cognition and Emotion*, pp. 1–12, 2021.

[78] J. W. Schwartz, J. W. Engelberg, and H. Gouzoules, "Was that a scream? Listener agreement and major distinguishing acoustic features," *Journal of Nonverbal Behavior*, vol. 44, no. 2, pp. 233–252, 2020.

[79] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2You – the imperial toolkit for multimodal profiling by end-to-end learning," 2018. arXiv:1802.01115.

[80] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of MM '13, 21st ACM International Conference on Multimedia*, (Barcelona, Spain), pp. 835–838, ACM, 2013.

[81] M. Schmitt and B. Schuller, "openXBOW – introducing the passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, 2017.

[82] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[83] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of ICASSP 2017, IEEE International Conference on Acoustics, Speech and Signal Processing*, (New Orleans, LA, USA), pp. 776–780, IEEE, 2017.

[84] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," 2016. arXiv:1609.08675.

[85] 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 *Proceedings of INTERSPEECH, 18th Annual Conference of the International Speech Communication Association*, (Stockholm, Sweden), pp. 3512–3516, ISCA, 2017.

[86] A. Paiva, I. Machado, and C. Martinho, "Enriching pedagogical agents with emotional behaviour – the case of Vincent," in *Proceedings of AI-ED '99, Workshop on Animated and Personified Pedagogical Agents*, (Le Mans, France), pp. 47–55, 1999.

[87] N. Yoon and H.-K. Lee, "AI recommendation service acceptance: Assessing the effects of perceived empathy and need for cognition," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 5, pp. 1912–1928, 2021.

[88] C. Pelau, D.-C. Dabija, and I. Ene, "What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic

characteristics in the acceptance of artificial intelligence in the service industry," *Computers in Human Behavior*, vol. 122, 2021. Art. no. 106855.

[89] B. Goertzel and C. Pennachin, *Artificial General Intelligence*. Springer, Berlin, Heidelberg, 2007.

[90] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[91] C. Montemayor, J. Halpern, and A. Fairweather, "In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare," *AI & Society*, 2021.

[92] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[93] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[94] J. A. Harrigan and K. T. Taing, "Fooled by a smile: Detecting anxiety in others," *Journal of Nonverbal Behavior*, vol. 21, no. 3, pp. 203–221, 1997.

[95] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[96] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[97] K. Stanisławski, J. Cieciuch, and W. Strus, "Ellipse rather than a circumplex: A systematic test of various circumplexes of emotions," *Personality and Individual Differences*, vol. 181, 2021. Art. no. 111052.

[98] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive Behavioural Systems* (A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Müller, eds.), pp. 144–157, Springer, Berlin, Heidelberg, 2012.

[99] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-Conversation corpus," in *Proceedings of INTERSPEECH, 21st Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 1823–1827, ISCA, 2020.

[100] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. W. Schuller, K. Star, E. Hajiyev, and M. Pantic, "SEWA DB: A rich database

for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, 2021.

[101] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 341–383, 2019.

[102] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, 2018. Art. no. e0196391.

[103] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, vol. 1, (Melbourne, Australia), pp. 2236–2246, 2018.

[104] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proceedings of 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (Shanghai, China), IEEE, 2013.

[105] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[106] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of INTERSPEECH, 6th Annual Conference of the International Speech Communication Association*, (Lisbon, Portugal), pp. 1517–1520, ISCA, 2005.

[107] M. Gerczuk, S. Amiriparian, S. Ottl, and B. Schuller, "EmoNet: A transfer learning framework for multi-corpus speech emotion recognition," 2021. arXiv:2103.08310.

[108] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," 2017. arXiv:1705.08045.

[109] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, 2021. Art. no. 1249.

[110] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proceedings of INTERSPEECH, 20th Annual Conference of the International Speech Communication Association*, (Graz, Austria), pp. 2803–2807, ISCA, 2019.

[111] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting," *Journal of Neuropsychobiology*, vol. 28, no. 1–2, pp. 76–81, 1993.

[112] A. Baird, S. Amiriparian, N. Cummins, S. Sturmbauer, J. Janson, E.-M. Messner, H. Baumeister, N. Rohleder, and B. Schuller, "Using speech to predict sequentially measured cortisol levels during a Trier Social Stress Test," in *Proceedings of INTERSPEECH, 20th Annual Conference of the International Speech Communication Association*, (Graz, Austria), pp. 534–538, ISCA, 2019.

[113] E. Childs, A. Dlugos, and H. De Wit, "Cardiovascular, hormonal, and emotional responses to the TSST in relation to sex and menstrual cycle phase," *Psychophysiology*, vol. 47, no. 3, pp. 550–559, 2010.

[114] R. Miller, F. Plessow, M. Rauh, M. Gröschl, and C. Kirschbaum, "Comparison of salivary cortisol as measured by different immunoassays and tandem mass spectrometry," *Psychoneuroendocrinology*, vol. 38, no. 1, pp. 50–57, 2013.

[115] A. T. Beck, N. Epstein, G. Brown, and R. A. Steer, "An inventory for measuring clinical anxiety: Psychometric properties," *Journal of Consulting and Clinical Psychology*, vol. 56, no. 6, pp. 893–897, 1988.

[116] J. Russel, "Core affect and the psychological construction of emotions," *Psychological Review*, vol. 110, no. 1, pp. 145–172, 2003.

[117] 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.

[118] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019. arXiv:1904.08779.

[119] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "How many labellers? Modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody," in *Second Language Studies: Acquisition, Learning, Education and Technology*, (Tokyo, Japan), 2010.

[120] A. Hadinejad, B. D. Moyle, A. Kralj, and N. Scott, "Physiological and self-report methods to the measurement of emotion in tourism," *Tourism Recreation Research*, vol. 44, no. 4, pp. 466–478, 2019.

[121] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[122] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[123] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proceedings of ASRU 2005, IEEE Workshop on Automatic Speech Recognition and Understanding*, (Cancun, Mexico), pp. 381–385, IEEE, 2005.

[124] S. Marioryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Proceedings of ACII 2013, Humaine Association Conference on Affective Computing and Intelligent Interaction*, (Geneva, Switzerland), pp. 85–90, IEEE, 2013.

[125] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," *Proceedings of NIPS 2009, Advances in Neural Information Processing Systems*, vol. 22, pp. 2286–2294, 2009.

[126] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, NY, USA: Wiley, 1958.

[127] F. Zhou and F. De la Torre, "Generalized canonical time warping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 279–294, 2015.

[128] K. R. Rao, D. N. Kim, and J. J. Hwang, *Fast Fourier Transform: Algorithms and Applications*. Springer, Berlin, Heidelberg, 2010.

[129] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[130] W. J. Hess, "Pitch determination of speech signals – a survey," in *Spoken Language Generation and Understanding*, pp. 263–278, Springer, Dordrecht, 1980.

[131] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[132] G. Muhammad, "Extended average magnitude difference function based pitch detection," *The International Arab Journal of Information Technology*, vol. 8, no. 2, pp. 197–203, 2011.

[133] K. Tom, I. R. Titze, E. A. Hoffman, and B. H. Story, "Three-dimensional vocal tract imaging and formant structure: Varying vocal register, pitch, and loudness," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 742–747, 2001.

[134] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[135] B. B. Bauer, E. L. Torick, and R. G. Allen, "The measurement of loudness level," *The Journal of the Acoustical Society of America*, vol. 50, no. 2A, pp. 405–414, 1971.

[136] K. R. Scherer, *Depression and Expressive Behavior :Vocal assessment of affective disorders*. London, UK: Lawrence Erlbaum Associatees, 1987.

[137] F. Ringeval, M. Chetouani, and B. Schuller, "Novel metrics of speech rhythm for the assessment of emotion," in *Proceedings of INTERSPEECH, 13th Annual Conference of the International Speech Communication Association*, (Portland, OR, USA), pp. 2763–2766, ISCA, 2012.

[138] S. J. Mozziconacci and D. J. Hermes, "Expression of emotion and attitude through temporal speech variations," in *Proceedings of ICSLP 2000, Sixth International Conference on Spoken Language Processing*, (Beijing, China), 2000.

[139] J. Fernandes, F. Teixeira, V. Guedes, A. Junior, and J. P. Teixeira, "Harmonic to noise ratio measurement-selection of window and length," *Procedia Computer Science*, vol. 138, pp. 280–285, 2018.

[140] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[141] P. Mermelstein, *Pattern recognition and artificial intelligence: Distance measures for speech recognition, psychological and instrumental*. Cambridge, MA, US: Academic Press Inc, 1976.

[142] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, and A. Stauffer, "Survey and evaluation of acoustic features for speaker recognition," in *Proceedings of ICASSP 2011, IEEE International Conference on Acoustics, Speech and Signal Processing*, (Prague, Czechia), pp. 5444–5447, IEEE, 2011.

[143] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proceedings of SCIPY 2015, 14th Python in Science Conference*, vol. 8, (Austin, Texas), pp. 18–25, 2015.

[144] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of MM '10, 18th ACM International Conference on Multimedia*, (Firenze, Italy), pp. 1459–1462, ACM, 2010.

[145] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Proceedings of ASMMC-MMAC'18, Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and 1st Multi-Modal Affective Computing of Large-Scale Multimedia Data*, (Seoul, Korea), pp. 27–33, 2018.

[146] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta oto-laryngologica*, vol. 90, no. 1-6, pp. 441–451, 1980.

[147] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[148] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of NIPS 2012, Advances in Neural Information Processing Systems*, vol. 25, (Lake Tahoe, NV, USA), pp. 1097–1105, 2012.

[149] S. Amiriparian, *Deep representation learning techniques for audio signal processing*. PhD thesis, Technische Universität München, Germany, 2019.

[150] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. arXiv:1409.1556.

[151] M. Ebersbach, R. Herms, and M. Eibl, "Fusion methods for ICD10 code classification of death certificates in multilingual corpora," in *18th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2017*, (Dublin, Ireland), 2017.

[152] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[153] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[154] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[155] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA, USA: Apress, 2015.

[156] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of NIPS 2014, Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), vol. 27, (Montreal, Canada), pp. 2672–2680, 2014.

[157] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. arXiv:1412.6980.

[158] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[159] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Proceedings of NIPS 2000, Advances in Neural Information Processing Systems*, vol. 13, (Denver, CO, USA), pp. 402–408, 2001.

[160] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, "Recurrent neural networks for language understanding," in *Proceedings of INTERSPEECH, 14th Annual Conference of the International Speech Communication Association*, (Lyon, France), pp. 2524–2528, ISCA, 2013.

[161] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proceedings of INTERSPEECH, 17th Annual Conference of the International Speech Communication Association*, (San Francisco, CA, USA), pp. 3603–3607, ISCA, 2016.

[162] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[163] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014. arXiv:1409.0473.

[164] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," 2016. arXiv:1601.06733.

[165] D. Barić, P. Fumić, D. Horvatić, and T. Lipic, "Benchmarking attention-based interpretability of deep learning in multivariate time series predictions," *Entropy*, vol. 23, no. 2, 2021. Art. no. 143.

[166] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. Van Essen, A. A. Awwal, and V. K. Asari, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, no. 3, 2019. Art. no. 292.

[167] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of NIPS 2017, Advances in Neural Information Processing Systems*, vol. 30, (Long Beach, CA, USA), pp. 4077–4087, 2017.

[168] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, and M. Sun, "FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation," 2018. arXiv:1810.10147.

[169] J. Pons, J. Serrà, and X. Serra, "Training neural audio classifiers with few data," in *Proceedings of ICASSP 2019, IEEE International Conference on Acoustics, Speech and Signal Processing*, (Brighton, UK), pp. 16–20, IEEE, 2019.

[170] A. Baird, S. Amiriparian, and B. Schuller, "Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation," in *Proceedings of MMSP 2019, 21st International Workshop on Multimedia Signal Processing*, (Kuala Lumpur, Malaysia), IEEE, 2019.

[171] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models," 2021. arXiv:2103.04922.

[172] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," 2018. arXiv:1802.04208.

[173] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015. arXiv:1511.06434.

[174] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016. arXiv:1609.03499.

[175] A. Borji, "Pros and cons of GAN evaluation measures: New developments," 2021. arXiv:2103.09396.

[176] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proceedings of NIPS 2016, Advances in Neural Information Processing Systems*, vol. 29, (Barcelona, Spain), pp. 2234–2242, 2016.

[177] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proceedings of NIPS 2017, Advances in Neural Information Processing Systems*, vol. 30, (Long Beach, CA, USA), 2017.

[178] S. Barua, X. Ma, S. M. Erfani, M. E. Houle, and J. Bailey, "Quality evaluation of GANs using cross local intrinsic dimensionality," 2019. arXiv:1905.00643.

[179] H. Yaribeygi, Y. Panahi, H. Sahraei, T. P. Johnston, and A. Sahebkar, "The impact of stress on body function: A review," *EXCLI Journal: Experimental and Clinical Sciences*, vol. 16, pp. 1057–1072, 2017.

[180] J. Thijssen, J. Van den Berg, H. Adlercreutz, A. Gijzen, F. De Jong, J. Meijer, and A. Moolenaar, "The determination of cortisol in human plasma: Evaluation and comparison of seven assays," *Clinica Chimica Acta*, vol. 100, no. 1, pp. 39–46, 1980.

[181] J. M. Turner-Cobb, "Psychological and stress hormone correlates in early life: A key to HPA-axis dysregulation and normalisation," *Stress*, vol. 8, no. 1, pp. 47–57, 2005.

[182] G. Hagerer, V. Pandit, F. Eyben, and B. Schuller, "Enhancing LSTM RNN-based speech overlap detection by artificially mixed data," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, (Erlangen, Germany), 2017.

[183] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[184] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," in *Proceedings of MuSe'20, 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, (Seattle, WA, USA), pp. 27–34, ACM, 2020.

[185] A. E.-D. Mousa and B. W. Schuller, "Deep bidirectional long short-term memory recurrent neural networks for grapheme-to-phoneme conversion utilizing complex many-to-many alignments," in *Proceedings of INTERSPEECH, 17th Annual Conference of the International Speech Communication Association*, (San Francisco, CA, USA), pp. 2836–2840, ISCA, 2016.

[186] G. G. Berntson and J. T. Cacioppo, "Heart rate variability: Stress and psychiatric conditions," in *Dynamic Electrocardiography*, pp. 57–64, Oxford, UK: Blackwell Publishing, 2004.

[187] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel, "Influence of mental stress on heart rate and heart rate variability," in *4th European Conference of the International Federation for Medical and Biological Engineering*, (Antwerp, Belgium), pp. 1366–1369, 2009.

[188] D. S. Goldstein, "Stress-induced activation of the sympathetic nervous system," *Bailliere's Clinical Endocrinology and Metabolism*, vol. 1, no. 2, pp. 253–278, 1987.

[189] S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. Berger, and R. J. Cohen, "Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control," *Science*, vol. 213, no. 4504, pp. 220–222, 1981.

[190] R. F. Orlikoff and R. Baken, "The effect of the heartbeat on vocal fundamental frequency perturbation," *Journal of Speech, Language, and Hearing Research*, vol. 32, no. 3, pp. 576–582, 1989.

[191] B. Schuller, F. Friedmann, and F. Eyben, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *Proceedings of*

*ICASSP 2013, International Conference on Acoustics, Speech and Signal Processing*, (Vancouver, Canada), pp. 7219–7223, IEEE, 2013.

[192] A. Jati, P. G. Williams, B. Baucom, and P. Georgiou, "Towards predicting physiology from speech during stressful conversations: Heart rate and respiratory sinus arrhythmia," in *Proceedings of ICASSP 2018, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4944–4948, IEEE, 2018.

[193] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, 2010.

[194] K. McKenzie, A. Murray, and T. Booth, "Do urban environments increase the risk of anxiety, depression and psychosis? An epidemiological study," *Journal of Affective Disorders*, vol. 150, no. 3, pp. 1019–1024, 2013.

[195] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Arlington, VA, USA: APA, 2013.

[196] E. A. Holmes, R. C. O'Connor, V. H. Perry, I. Tracey, S. Wessely, L. Arseneault, C. Ballard, H. Christensen, R. Cohen Silver, I. Everall, T. Ford, A. John, T. Kabir, K. King, I. Madan, S. Michie, A. K. Przybylski, R. Shafran, A. Sweeney, C. M. Worthman, L. Yardley, K. Cowan, C. Cope, M. Hotopf, and E. Bullmore, "Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science," *The Lancet Psychiatry*, vol. 7, no. 6, pp. 547–560, 2020.

[197] M. Shevlin, O. McBride, J. Murphy, J. G. Miller, T. K. Hartman, L. Levita, L. Mason, A. P. Martinez, R. McKay, T. V. Stocks, K. M. Bennett, P. Hyland, T. Karatzias, and R. P. Bentall, "Anxiety, depression, traumatic stress, and COVID-19 related anxiety in the UK general population during the COVID-19 pandemic," *BJPsych Open*, vol. 6, no. 6, 2020. Art. no. E125.

[198] J. Peat and B. Barton, *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*. Malden, MA, USA: Blackwell Publishing Ltd, 2005.

[199] P. Laukka, C. Linnman, F. Åhs, A. Pissiota, Ö. Frans, V. Faria, Å. Michelgård, L. Appel, M. Fredrikson, and T. Furmark, "In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech," *Journal of Nonverbal Behavior*, vol. 32, no. 4, pp. 195–214, 2008.

[200] K. Kvaal, I. Ulstein, I. H. Nordhus, and K. Engedal, "The Spielberger State-Trait Anxiety Inventory (STAI): The state scale in detecting mental disorders in geriatric patients," *International Journal of Geriatric Psychiatry: A Journal of the Psychiatry of Late Life and Allied Sciences*, vol. 20, no. 7, pp. 629–634, 2005.

[201] M. Cook, "Anxiety, speech disturbances and speech rate," *British Journal of Social and Clinical Psychology*, vol. 8, no. 1, pp. 13–21, 1969.

[202] J. A. Harrigan and D. M. O'Connell, "How do you look when feeling anxious? Facial displays of anxiety," *Personality and Individual Differences*, vol. 21, no. 2, pp. 205–212, 1996.

[203] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception & Psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[204] G. D. Bodie, "A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety," *Communication Education*, vol. 59, no. 1, pp. 70–105, 2010.

[205] T. J. Davis, M. Morris, and M. M. Drake, "The moderation effect of mindfulness on the relationship between adult attachment and wellbeing," *Personality and Individual Differences*, vol. 96, pp. 115–121, 2016.

[206] N. S. Schutte and J. M. Malouff, "Emotional intelligence mediates the relationship between mindfulness and subjective well-being," *Personality and Individual Differences*, vol. 50, no. 7, pp. 1116–1119, 2011.

[207] A. N. Niles and M. G. Craske, "Incidental emotion regulation deficits in public speaking anxiety," *Cognitive Therapy and Research*, vol. 43, no. 2, pp. 419–426, 2019.

[208] B. W. Schuller and A. M. Batliner, *Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, 1988.

[209] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.

[210] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of ICASSP 2017,*

*IEEE International Conference on Acoustics, Speech and Signal Processing*, (New Orleans, LA, USA), pp. 2227–2231, IEEE, 2017.

[211] L. Moussu and E. Llurda, "Non-native English-speaking English language teachers: History and research," *Language Teaching*, vol. 41, no. 3, pp. 315–348, 2008.

[212] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of ICASSP 2016, International Conference on Acoustics, Speech and Signal Processing*, (Shanghai, China), pp. 5200–5204, 2016.

[213] A. S. Thompson and L. K. Sylvén, ""does english make you nervous?' anxiety profiles of clil and non-clil students in sweden," *Apples-Journal of Applied Language Studies*, vol. 9, no. 2, pp. 1–23, 2015.

[214] S. Duijndam, A. Karreman, J. Denollet, and N. Kupper, "Physiological and emotional responses to evaluative stress in socially inhibited young adults," *Biological Psychology*, vol. 149, 2020. Art. no. 107811.

[215] J. Campbell and U. Ehlert, "Acute psychosocial stress: Does the emotional stress response correspond with physiological responses?," *Psychoneuroendocrinology*, vol. 37, no. 8, pp. 1111–1134, 2012.

[216] L. Bernardi, J. Wdowczyk-Szulc, C. Valenti, S. Castoldi, C. Passino, G. Spadacini, and P. Sleight, "Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability," *Journal of the American College of Cardiology*, vol. 35, no. 6, pp. 1462–1469, 2000.

[217] N. S. Eckland, T. M. Leyro, W. B. Mendes, and R. J. Thompson, "The role of physiology and voice in emotion perception during social stress," *Journal of Nonverbal Behavior*, vol. 43, no. 4, pp. 493–511, 2019.

[218] D. Caruelle, A. Gustafsson, P. Shams, and L. Lervik-Olsen, "The use of electrodermal activity (EDA) measurement to understand consumer emotions – a literature review and a call for action," *Journal of Business Research*, vol. 104, pp. 146–160, 2019.

[219] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen, "Correlation between heart rate, electrodermal activity and player experience in first-person shooter games," in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, (Los Angeles, CA, USA), pp. 49–54, 2010.

[220] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proceedings of ICMI '20, International Conference on Multimodal Interaction*, pp. 784–789, ACM, 2020.

[221] A. M. Goberman, S. Hughes, and T. Haydock, "Acoustic characteristics of public speaking: Anxiety and practice effects," *Speech Communication*, vol. 53, no. 6, pp. 867–876, 2011.

[222] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3705–3722, 2019.

[223] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[224] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "StarGAN for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *Proceedings of ICASSP 2020, International Conference on Acoustics, Speech and Signal Processing*, (Barcelona, Spain), pp. 3502–3506, IEEE, 2020.

[225] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation," 2019. arXiv:1904.11685.

[226] M. A. Cohen, T. S. Horowitz, and J. M. Wolfe, "Auditory recognition memory is inferior to visual recognition memory," *Proceedings of the National Academy of Sciences*, vol. 106, no. 14, pp. 6008–6010, 2009.

[227] A. Batliner, S. Hantke, and B. W. Schuller, "Ethics and good practice in computational paralinguistics," *IEEE Transactions on Affective Computing*, 2020.

[228] K. Johnson, F. Pasquale, and J. Chapman, "Artificial intelligence, machine learning, and bias in finance: Toward responsible innovation," *Fordham Law Review*, vol. 88, pp. 499–529, 2019.

[229] R. Srinivasan and K. Uchino, "Biases in generative art: A causal look from the lens of art history," in *Proceedings of ACM FAccT 2021, Conference on Fairness, Accountability, and Transparency*, pp. 41–51, 2021.

[230] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[231] W. Gao and H. Ai, "Face gender classification on consumer images in a multiethnic environment," in *Proceedings of International Conference on Advances in Biometrics*, (Alghero, Italy), pp. 169–178, 2009.

[232] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proceedings of ICCV '19, IEEE International Conference on Computer Vision*, (Seoul, Korea), pp. 5310–5319, IEEE, 2019.

[233] A. Koene, "Algorithmic bias: Addressing growing concerns," *IEEE Technology and Society Magazine*, vol. 36, no. 2, pp. 31–32, 2017.

[234] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, no. 3, pp. 231–237, 2019.

[235] C. Sueur, J.-L. Deneubourg, and O. Petit, "From social network (centralized vs. decentralized) to collective decision-making (unshared vs. shared consensus)," *PLOS ONE*, vol. 7, no. 2, pp. 1–10, 2012. Art. no e32566.

[236] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[237] A. Baird, S. H. Jørgensen, E. Parada-Cabaleiro, N. Cummins, S. Hantke, and B. Schuller, "The perception of vocal traits in synthesized voices: Age, gender, and human likeness," *Journal of the Audio Engineering Society*, vol. 66, no. 4, pp. 277–285, 2018.

[238] Y. Wang and Q. Yao, "Few-shot learning: A survey," 2019. arXiv:1904.05046.

[239] R. Azad, A. R. Fayjie, C. Kauffman, I. B. Ayed, M. Pedersoli, and J. Dolz, "On the texture bias for few-shot CNN segmentation," 2020. arXiv:2003.04052.

[240] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, "General pitfalls of model-agnostic interpretation methods for machine learning models," 2020. arXiv preprint arXiv:2007.04131.

[241] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *Proceedings of ICASSP 2012, IEEE International Conference on Acoustics, Speech and Signal Processing*, (Kyoto, Japan), pp. 5117–5120, IEEE, 2012.

[242] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas, "A web crowdsourcing framework for transfer learning and personalized speech emotion recognition," *Machine Learning with Applications*, 2021. Art. no. 100132.

[243] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[244] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization," *American Psychologist*, vol. 74, no. 6, pp. 698–712, 2019.