REFERENCES

[1] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost. 2017. Pooling Acoustic and Lexical Features for the Prediction of Valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017)*. ACM, Glasgow, UK, 68–72.

[2] Z. Aldeneh and E. M. Provost. 2017. Using regional saliency for speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, New Orleans, Louisiana, 2741–2745.

[3] R. Alharbi, N. Vafaie, K. Liu, K. Moran, G. Ledford, A. Pfammatter, B. Spring, and N. Alshurafa. 2017. Investigating barriers and facilitators to wearable adherence in fine-grained eating detection. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, Kona, Hawaii, 407–412.

[4] S. Amiriparian, M. Gerczuk, S. Ottl, N.s Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller. 2017. Snore Sound Classification Using Image-based Deep Spectrum Features. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 3512–3516.

[5] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, M. Y. Prioleau, T.and Beh, M. Goel, T. Starner, and G. Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sep. 2017), 37:1–37:20.

[6] Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5, 2 (Mar. 1994), 157–166.

[7] R. Brueckner and B. Schulter. 2014. Social signal classification using deep blstm recurrent neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Florence, Italy, 4823–4827.

[8] R. Collobert, C. Puhrsch, and G. Synnaeve. 2016. Wav2Letter: an End-to-End ConvNet-based Speech Recognition System. *CoRR* abs/1609.03193 (2016).

[9] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller. 2017. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, Mountain View, California, 478–484.

[10] J.-B. Delbrouck and S. Dupont. 2017. Multimodal Compact Bilinear Pooling for Multimodal Neural Machine Translation. *CoRR* abs/1703.08084 (2017).

[11] R. Dobbs, C. Sawers, F. Thompson, J. Manyika, J. R. Woetzel, P. Child, S. McKenna, and A. Spatharou. 2014. Overcoming obesity: an initial economic analysis. https://goo.gl/6R7kz2. Accessed: 31-05-2018.

[12] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps. 2017. Bidirectional Modelling for Short Duration Language Identification. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 2809–2813.

[13] J. M. Fontana, M. Farooq, and E. Sazonov. 2014. Automatic Ingestion Monitor: A Novel Wearable Device for Monitoring of Ingestive Behavior. *IEEE Transactions on Biomedical Engineering* 61, 6 (June 2014), 1772–1779.

[14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *CoRR* abs/1606.01847 (2016).

[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. 2016. Compact Bilinear Pooling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, Nevada, 317–326.

[16] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen. 2018. Recent advances in convolutional neural networks. *Pattern Recognition* 77 (2018), 354–377.

[17] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller. 2016. Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals. *Image and Vision Computing, Special Issue on Multimodal Sentiment Analysis and Mining in the Wild* 65 (Sep. 2016), 76–86.

[18] S. Hantke, M. Schmitt, P. Tzirakis, and B. Schuller. 2018. EAT – The ICMI 2018 Eating Analysis and Tracking Challenge. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. ACM, Boulder, Colorado. 5 pages.

[19] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. El-Desoky Mousa, and B. Schuller. 2016. I Hear You Eat and Speak: Automatic Recognition of Eating Condition and Food Types, Use-Cases, and Impact on ASR Performance. *PLoS ONE* 11, 5 (May 2016), 1–24.

[20] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[21] C. W. Huang and S. S. Narayanan. 2017. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Hong Kong, P. R. China, 583–588.

[22] H. Kaya, A. A. Karpov, and A. A. Salah. 2015. Fisher vectors with cascaded normalization for paralinguistic analysis. In *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. ISCA, Dresden, Germany, 909–913.

[23] J.-H. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B.-T. Zhang. 2016. Hadamard Product for Low-rank Bilinear Pooling. *CoRR* abs/1610.04325 (2016).

[24] S. Kong and C. Fowlkes. 2017. Low-Rank Bilinear Pooling for Fine-Grained Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, Hawaii, 365–374.

[25] D. Le, Z. Aldeneh, and E. Mower Provost. 2017. Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 1108–1112.

[26] W. Lim, D. Jang, and T. Lee. 2016. Speech emotion recognition using convolutional and Recurrent Neural Networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, Jeju, South Korea, 1–4.

[27] E. A. Lin, G. M. Barlow, and R. Mathur. 2015. *The Health Burden of Obesity*. Springer New York, New York, NY, 19–42.

[28] V. Liptchinsky, G. Synnaeve, and R. Collobert. 2017. Letter-Based Speech Recognition with Gated ConvNets. *CoRR* abs/1712.09444 (2017).

[29] H. Liu, H. Ning, Q. Mu, Y. Zheng, J. Zeng, L. T. Yang, R. Huang, and J. Ma. 2017. A review of the smart world. *Future Generation Computer Systems* (2017). 14 pages, in press.

[30] B. Milde and C. Biemann. 2015. Using representation learning and out-of-domain data for a paralinguistic speech task. In *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. ISCA, Dresden, Germany, 904–908.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia, 5206–5210.

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

[33] R. Pham, N.and Pagh. 2013. Fast and Scalable Polynomial Kernels via Explicit Feature Maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, Chicago, Illinois, 239–247.

[34] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. 2015. The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition. In *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. ISCA, Dresden, Germany, 478–482.

[35] J. B. Tenenbaum and W. T. Freeman. 2000. Separating Style and Content with Bilinear Models. *Neural Computation* 12, 6 (June 2000), 1247–1283.

[36] M. Tremmel, U.-G. Gerdtham, P. M. Nilsson, and S. Saha. 2017. Economic Burden of Obesity: A Systematic Literature Review. *International Journal of Environmental Research and Public Health* 14, 4 (2017). Article Number 435.

[37] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou. 2016. Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network. In *Proceedings 41st IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2016*. IEEE, Shanghai, P. R. China, 5200–5204.

[38] World Health Organization (WHO). 2018. Obesity and Overweight. http://www.who.int/mediacentre/factsheets/fs311/en/. Accessed: 26-03-2018.

[39] Z. Yu, J. Yu, J. Fan, and D. Tao. 2017. Multi-Modal Factorized Bilinear Pooling With Co-Attention Learning for Visual Question Answering. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 1821–1830.

[40] S. Zhang, S. Zhang, T. Huang, and W. Gao. 2018. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia* 20, 6 (June 2018), 1576–1590.