REFERENCES

[1] Sumair Aziz, Muhammad Awais, Talha Akram, Muhammad Umar, Khursheed Khursheed, and Musaed Alhussein. 2019. Automatic Scene Recognition through Acoustic Classification for Behavioral Robotics. *Electronics* 8 (04 2019). https://doi.org/10.3390/electronics8050483

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014).

[3] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. 2005. A Review of Audio Fingerprinting. *J. VLSI Signal Process. Syst.* 41, 3 (Nov. 2005), 271–284.

[4] Gianmarco Cerutti, Rahul Prasad, Alessio Brutti, and Elisabetta Farella. 2019. Neural Network Distillation on IoT Platforms for Sound Event Detection. (08 2019).

[5] Shizhe Chen, Jia Chen, Qin Jin, and Alexander Hauptmann. 2018. Class-aware Self-Attention for Audio Event Recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (Yokohama, Japan) *(ICMR '18)*. ACM, New York, NY, USA, 28–36. https://doi.org/10.1145/3206025.3206067

[6] Xuerui Dai and Xueye Wei. 2018. HybridNet: A fast vehicle detection system for autonomous driving. *Signal Processing: Image Communication* 70 (09 2018). https://doi.org/10.1016/j.image.2018.09.002

[7] Jinxi Guo, Ning Xu, Li-Jia Li, and Abeer Alwan. 2017. Attention Based CLDNNs for Short-Duration Acoustic Scene Classification. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. 469–473. http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0440.html

[8] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J. F. Serignat. 2006. Information Extraction from Sound for Medical Telemonitoring. *Trans. Info. Tech. Biomed.* 10, 2 (April 2006).

[9] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel P. W. Ellis, Shawn Hershey, Jiayang Liu, R. Channing Moore, and Rif A. Saurous. 2017. Unsupervised Learning of Semantic Audio Representations. *CoRR* abs/1711.02209 (2017). arXiv:1711.02209 http://arxiv.org/abs/1711.02209

[10] Alfred Mertins and Dr Alfred Mertins. 1999. *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications*. John Wiley & Sons, Inc., New York, NY, USA.

[11] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT Database for Acoustic Scene Classification and Sound Event Detection. In *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*. Budapest, Hungary.

[12] Arthur William Moore and James W. Jorgenson. 1993. Median filtering for removal of low-frequency background drift. *Analytical chemistry* 65 2 (1993), 188–91.

[13] Nobutaka Ono, Kenichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka, and Shigeki Sagayama. 2008. Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. (01 2008).

[14] Huy Phan, Oliver Y. Chén, Lam Dang Pham, Philipp Koch, Maarten De Vos, Ian Vince McLoughlin, and Alfred Mertins. 2019. Spatio-Temporal Attention Pooling for Audio Scene Classification. *CoRR* abs/1904.03543 (2019). arXiv:1904.03543 http://arxiv.org/abs/1904.03543

[15] Huy Phan, Philipp Koch, Fabrice Katzberg, Marco Maaß, Radoslaw Mazur, and Alfred Mertins. 2017. Audio Scene Classification with Deep Recurrent Neural Networks. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. 3043–3047. http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0101.html

[16] Aref Pour, Mohammad Asgari, and Mohammad Hasanabadi. 2014. Gammatonegram based speaker identification. *Proceedings of the 4th International Conference on Computer and Knowledge Engineering, ICCKE 2014*, 52–55.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*.

[19] Nicolas Turpault, Romain Serizel, and Emmanuel Vincent. 2019. Semi-supervised Triplet Loss Based Learning of Ambient Audio Embeddings. 760–764. https://doi.org/10.1109/ICASSP.2019.8683774

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., USA, 6000–6010. http://dl.acm.org/citation.cfm?id=3295222.3295349

[21] Wired 2017. Wired. https://www.wired.com/story/driverless-cars-need-ears-as-well-as-eyes/.

[22] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. 2017. Sampling Matters in Deep Embedding Learning. *CoRR* abs/1706.07567 (2017). arXiv:1706.07567 http://arxiv.org/abs/1706.07567

[23] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. 2019. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. In *Sensors*.

[24] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. 2001. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology* 16, 6 (01 Nov 2001), 582–589. https://doi.org/10.1007/BF02943243