13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

Acoustic Traffic Monitoring with MEMS Microphones, Concatenated log-Bark Spectrograms and CNN

István Pintér[a], **Lóránt Kovács**[a]

[a]John von Neumann University, GAMF Faculty of Engineering and Computer Science
pinter.istvan@nje.hu

Abstract

Recent advances in info-communication technologies highlight the concept of smart cities. One objective is the optimal control of urban and suburban vehicle traffic, which requires traffic monitoring. It's main problems are the following: 1) detection and counting of passing vehicles, 2) determination of vehicle's category, 3) determination of motion's direction, and 4) vehicle's speed estimation. The real-time data necessary for traffic monitoring are supplied by appropriate sensors. Inductive loops and pneumatic pipes are highly reliable ones for solving problem 1) and 3). Unfortunately, these are not suitable for real-time traffic monitoring: inductive loops must be built into the road during road-building, and pneumatic pipes must be laid across the road, thus affecting the traffic itself. Another family of sensors are based on electromagnetic operating principles (e.g. radar, lidar, and cameras). However, their - otherwise excellent - detection accuracy is highly dependent on meteorological conditions and time of day. Besides, the high cost of installation and maintenance also calls for another or additional sensor-domain, namely acoustic sensors. This leads to the concept of Acoustic Traffic Monitoring (ATM). Most frequently, pressure-sensing microphones are used, and a promising goal is developing sensors with low-cost, medium-quality omnidirectional MEMS (Micro-ElectroMechanical Systems) microphones. The main efforts in this sensor-development field can be summarized as: I) minimizing the cost of sensors, II) decreasing the installation and maintenance costs, (III) decreasing the electrical power consumption of the sensor and signal processing units, (IV) optimizing the suitable sensor network architecture for covering a large geographic area with many

cooperating sensor units, and (V) guaranteeing the continuous operation even in adverse weather conditions. We present our results in algorithm-development for ATM problems, using MEMS microphone-signals. Our aim was that the algorithms have to be realised in the sensor itself, thus we followed the edge-AI system-concept. Several freely available audio datasets exist for this purpose. We used the IDMT-Traffic dataset, because it consists of two-channel (stereo) audio recordings both with high-quality sE8 and MEMS microphones. Therefore, the performance of MEMS-based algorithms can be compared with sE8-based ones. We considered [1] as reference, because the authors presented possible solutions for problems 1), 2), and 3) in case of sE8 microphones. In solving the problem 3) their neural network was a variant of MobileNet, adapted to the problem at hand. The input was a cross-correlogram, computed as a series of cross-correlations between the two microphone-channels. For problem 2) their classifier was a pretrained CNN (Convolutional Neural Network), which was tailored to the concrete problem using transfer learning. The input in this case was the so-called log-mel spectrogram. The log-mel spectrogram is computed from linear spectrogram using the mel-frequency scale and triangle-shaped, overlapped filters. The mel frequency-scale is a characteristics of human hearing, and its non-linear function is the result of function-fitting to experimental, psychoacoustical data [4]. However, this is not the only one in this respect; another is the Bark-scale [5]. Therefore, in a preliminary study - with the subject of detecting passing vehicles in background noise - we compared the effect of the mel/Bark scale on detection accuracy. Beyond that, in that comparison we used both triangular-shaped filters and raised cosine filters, too. The final result of the study was, that in case of sE8 microphone the best combination is mel-scale with raised-cosine filters, and in case of MEMS microphone the Bark-scale with raised cosine filters gave the best accuracy. Therefore, in this work we used the log-Bark spectrogram computed with raised cosine filters. In order to preserve the time-delay information between the two microphone channels, we concatenated the log-Bark spectrograms of the two channels ensuring the proper time-alignment. So, the concatenated log-Bark spectrogram was the result of feature extraction. It was then converted to grayscale image, which was the input of an image classifier. For the classifier we developed our own CNN-structure [2]. It is interesting to note, that we got finally the same structure when solving problem 2) and 3); only the parameter sets are different (the number of learnable parameters are 26011 in both cases). The vehicle-category determination was a 3-class classification problem (passenger car/commercial vehicle/no vehicle), and the overall accuracy achieved was 94.04±0.48%. The motion's direction determination was also a 3-class classification problem (from left/from right/no vehicle), the overall accuracy in this case was 97.70±0.64%. Problem 1) is a 2-class classification problem, and its confusion matrix can be derived from the 3-class one. The accuracy, F1-score and Matthew's Correlation Coefficient (MCC) [3] performance parameters were computed from the class-2 confusion matrix . These are (99.64%, 0.9965, 0.9928) in problem 2) and (99.75%, 0.9975, 0.9949) in problem 3). The accuracy of the latter is somewhat greater, so the problem 1) can also be solved

with the motion's direction classifier. Based on the above we concluded, that the MEMS-microphones could be used to develop edge-AI type sensors for the first three ATM problems using the concatenated log-Bark spectrograms converted to grayscale images as input to corresponding CNN-structure.

References

[1] J. Abesser, S. Gourishetti, et al.: *IDMT-Traffic: An Open Benchmark Dataset for Acoustic Traffic Monitoring Research*, in: Proceedings of 29th EUSIPCO European Signal Processing Conference, Dublin, Ireland, 2021, pp. 551–555, doi: 10.23919/EUSIPCO54536.2021.9616080.

[2] C. M. Bishop, H. Bishop: *Deep Learning, Foundations and Concepts*, Springer Cham, 2024, doi: 10.1007/978-3-031-45468-4.

[3] D. Chicco, G. Jurman: *The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification*, BioData Mining 16.4 (2023), pp. 1–23, doi: 10.1186/s13040-023-00322-4.

[4] S. S. Stevens, J. Volkmann, E. B. NewMan: *A scale for the measurement of the psychological magnitude pitch*, American. Journal of Psychology 8.3 (1937), pp. 185–190, doi: 10.1121/1.1915893.

[5] E. Zwicker: *Subdivision of the audible frequency range into critical bands (frequenzgruppen)*, Journal of the Acoustical Society of America 33.2 (1961), p. 248, doi: 10.1121/1.1908630.