Post-print of: Czyżewski A., Kurowski A., Zaporowski Sz.: Application of autoencoder to traffic noise analysis. Journal of the Acoustical Society of America. Vol. 146, iss. 4, 2958 (2019). DOI: https://doi.org/10.1121/1.5137275

Application of autoencoder to traffic noise analysis

Andrzej Czyzewski, Adam Kurowski, Szymon Zaporowski
Multimedia Systems Department, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Narutowicza 11/12, Gdansk 80-233, Poland;
ac@pg.edu.pl, adakurow@multimed.org, smck@multimed.org

The aim of an autoencoder neural network is to transform the input data into a lower-dimensional code and then to reconstruct the output from this representation. Applications of autoencoders to classifying sound events in the road traffic have not been found in the literature. The presented research aims to determine whether such an unsupervised learning method may be used for deploying classification algorithms applied to the automatic annotation of road traffic-related events based on noise analysis. Two-dimensional representation of traffic sounds based on 1D convolution was fed the core of autoencoder neural network, and after that classified with seven feed-forward classification subnetworks. Obtained results show that sound recordings can help determine the number of vehicles passing on the road. However, instead of being treated as independent, this method output should be combined with another source of data, e.g., video processing results or microwave radar data readings. Results of vehicle types classification and occupied lane obtained with the use of autoencoder are shown in the paper.

1. INTRODUCTION

Classification of vehicles based on the audio signal is a topic that appears in many works on traffic-related issues. Until now, the standard approach has been to use k-Nearest Neighbors algorithm, Support Vector Machine classifiers or simple neural networks[1]–[3]. Previously, solutions like Hidden Markov Model and their modifications were used. Spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding or low-dimension MPEG-7 descriptors were employed as parameters[1], [4]. Representation learning is a widely used technique, which can be used to parameterize audio signals such as acoustic scene ambient sound, music, speech or heart sounds [5], [6], [7]. The main difference between related works using autoencoder is feature extraction block (MFCC calculation) despite employing raw wave files as a feed [6]. Also previously mentioned approach using 1D convolution is based on MFCC feature maps [7]. In our research, we intended to perform unsupervised analysis of sounds gathered in the proximity of a road which may allow tasks such as classification of vehicle types or vehicle counting. These are examples of classification tasks providing crucial information required by intelligent transportation systems to perform tasks such as speed limit optimization. The novelty of our approach presented in the article is the use the auto encoder to classify vehicles and fact that this solution is characterized by the lack of parameterization of audio recordings due to the application of 1D convolutions in the context of parameterization of vehicle noise signals.

The proposed solution uses the Acoustic Vector Sensor (AVS) as an audio data source [8]. Such an acoustical probe application in the context of traffic classification can be very useful, for example by installing it inside an intelligent road sign that monitors the expressways constantly. Examples of such solutions are presented in Figure 1. There are several studies on traffic noise analysis based on AVS, including vehicle classification [2], [9]–[12]. The presented auto encoder may be one of the solutions used for intelligent road signs as a vehicle classification module or for speed measurement purposes employing AVS [11].

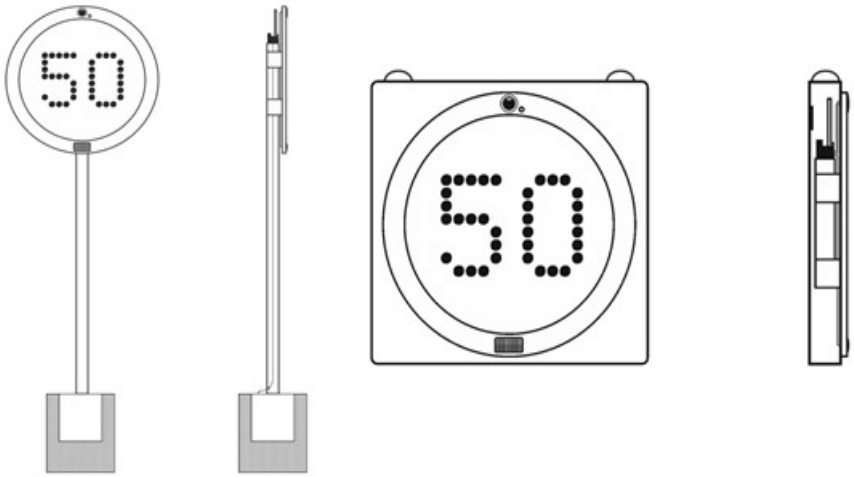

Figure 1. Standing and hanging smart road signs (patented utility models [13] [14]**)**

2. METHODOLOGY

The acoustic signal is gathered by a MEMS microphone-based acoustic probe (AVS) employing IvenSense INMP441 sensors. Moreover, a video signal is captured for the purpose of its use in the process of manual labeling of acoustic data. Both microphone and the camera can be seen with the measurement environment in Figure 2.

Figure 2. Hanging measurement module of an intelligent road sign in Lezno, Poland. At the bottom of the suspended box, marked with a red ellipse, microphone and camera is seen.

Seven types of labels are taken into consideration which describes both the type of the vehicle passing by at the given moment of time and the lanes being occupied at that moment. The labels for the vehicle type include car, bus, truck, motorcycle and van. For the lane identification, there are two labels – closer lane and far lane (relative to the position of the microphone). An example of such a labeled audio signal is shown in Figure 3.

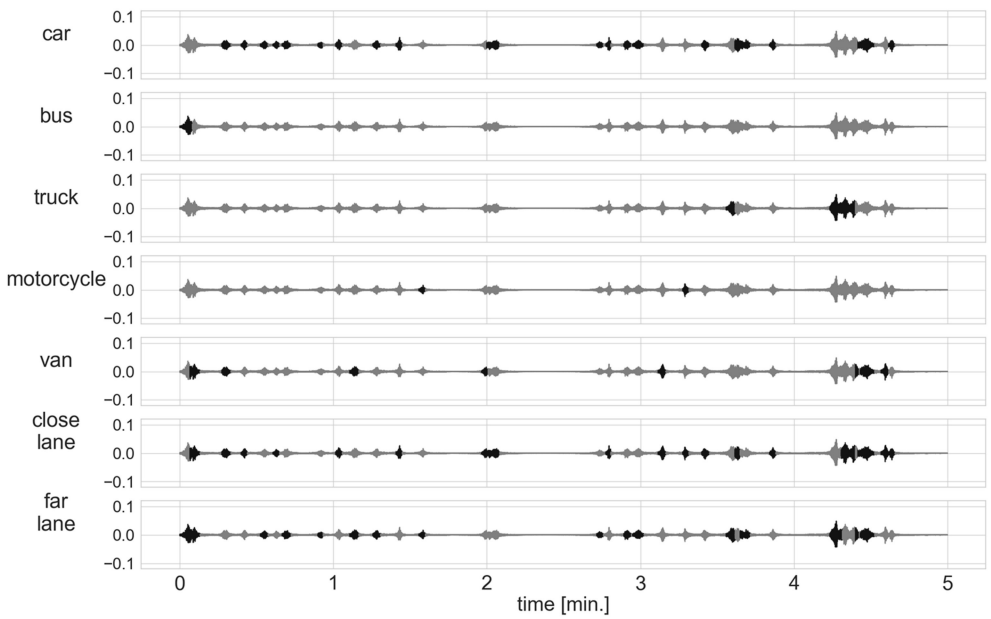

Figure 3. A visual depiction of labels assigned to audio segments from the manually annotated dataset with fragments marked in black. Each label represents a given vehicle type presence or presence of a vehicle on a given road lane.

Each frame can be associated with only one class, but it also can be connected to no class or more than one class if multiple vehicles are passing. If no class is assigned to the audio frame, then the frame contains only sounds related to the acoustic environment. Eight hours of unlabelled audio was used to train an autoencoder in the form of a type 1D (one dimension) convolutional neural network. For assessment of the quality of the unsupervised learning stage and for the training of the audio classifier in a supervised manner, an additional 2 hours of audio were manually labeled. Numbers of examples related to each class are shown in Table 1.

Table 1. The number of examples associated with each class in the labeled dataset. Silence (no class) means that no vehicle or lane-associated class were assigned to a given frame.

number of frames counted	silence (no class)	car	bus	truck	motorcycle	van	close lane	far lane
14160	9673	3441	103	613	96	431	2277	2414

Autoencoder neural network has the ability to perform unsupervised analysis of the structure of data fed to its input. The architecture employed in this study forces the network to represent each sample of data as a set of 64 channels, each consisting of 188 parameters. In total, each representation consists of $64 \cdot 188 = 12032$ coefficients. Such a representation of the audio frame is used by the decoding part of the network to recreate the original input signal. Decoder architecture is built in the same way as the encoding part of the network; however, it is a mirror reflection of the encoder structure. The task of the network during the training process is to minimize the reconstruction error. The encoding part of this network was used as the first part of 1D convolutional network which is performing classification. Two hours of audio recording were used for the training of the neural network. Frames of audio signals were 500 ms long. The sampling rate of the audio signal is 48 kHz. The general structure of the neural network used for the study is depicted in Figure 4.

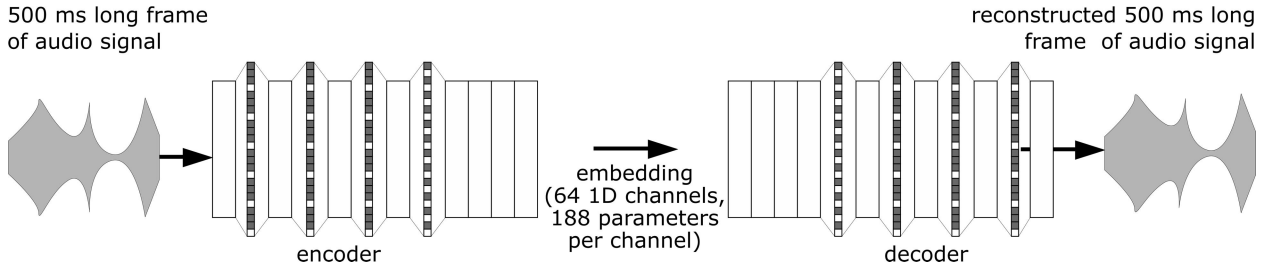

Figure 4. A general depiction of the unsupervised stage of pre-training

To interpret the output of the encoder, an additional set of sub-networks consisting of 7 feed-forward neural networks to perform classification tasks were added. Their inference is based upon an input from the encoder part of the autoencoder network. The general diagram of the proposed solution is presented in Figure 5. The output of the encoding part of the autoencoder is fed into 7 independent feed-forward neural networks. Each of the networks is associated with one of seven labels used for the study. The classification for each class is performed independently from each other, because a single frame can belong to multiple classes at the same time. The result of the classification is returned by the network in the form of seven two-element lone one-hot vectors.

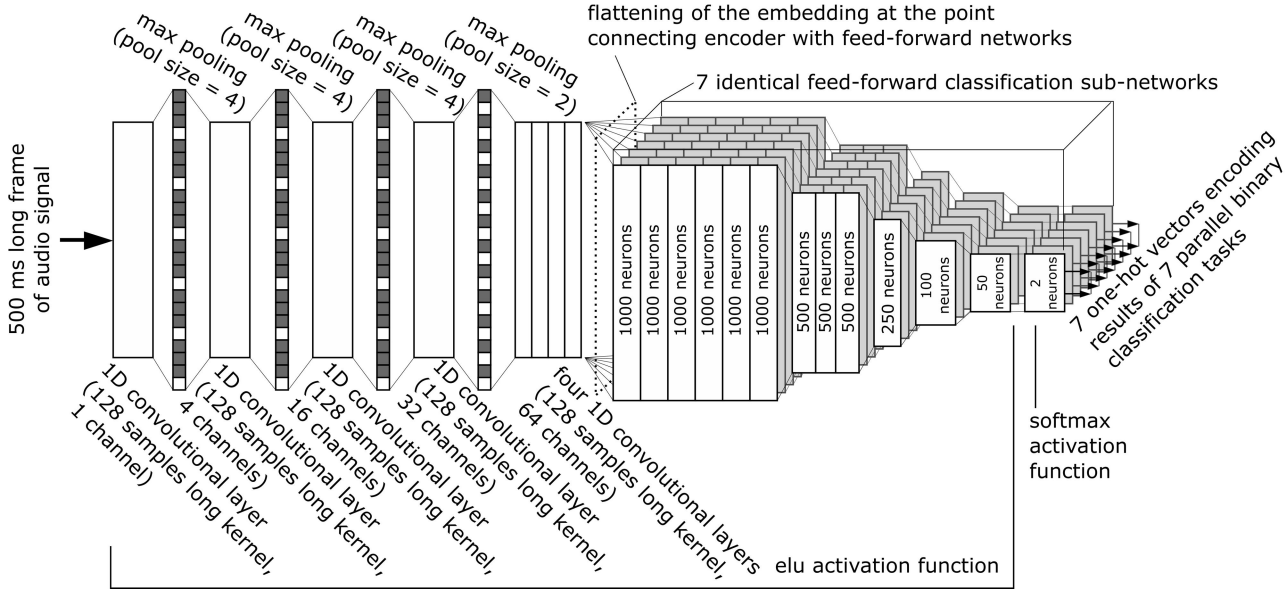

Figure 5. Structure of the classifier neural network employed in the supervised stage of the experiment

As a learning rate optimizer, the ADAM algorithm was employed in both the unsupervised and supervised learning-based stage of the experiment [15]. Learning rate in both cases was set to 10^{-4}.

3. RESULTS AND DISCUSSION

Further analyses were carried out for both the results of unsupervised and supervised learning processes. The unsupervised learning stage resulted in the assignment of frames of audio signals to points in a decision space consisting of 12032 dimensions. Each dimension represents one of the coefficients of the representation generated by the encoding part of the autoencoder neural network. An easy way of a quick assessment of the results of such a process is to employ PCA to reduce the dimensionality of the decision space obtained from the autoencoder and to display points in the form of clusters using only two first components of PCA, which contain the most significant values of the original set variance. Graphical depiction of the autoencoder-based clusterization calculated in this manner is presented in Figures 6 and 7. Each point depicted in Figures 6 and 7 is related to the frame belonging to a given class. Also, centroids of each class-related datasets are depicted in those figures, as its placement carries useful information about the possible achievable separation between

classes of sounds analyzed by the neural networks. The network itself had no information about the class of sound it processes. The result visible in Figures 6 and 7 was obtained by parameterization of frames obtained from a manually labeled fragment of the dataset, for which associated class labels were known and could be depicted in the image. Additionally, a so-called centroid of the joint set created by merging all single-class datasets shown in the figure is also presented. It may serve as a reference in calculation of the influence of the choice of points belonging to a certain class on the position of the cluster to which those data points belong.

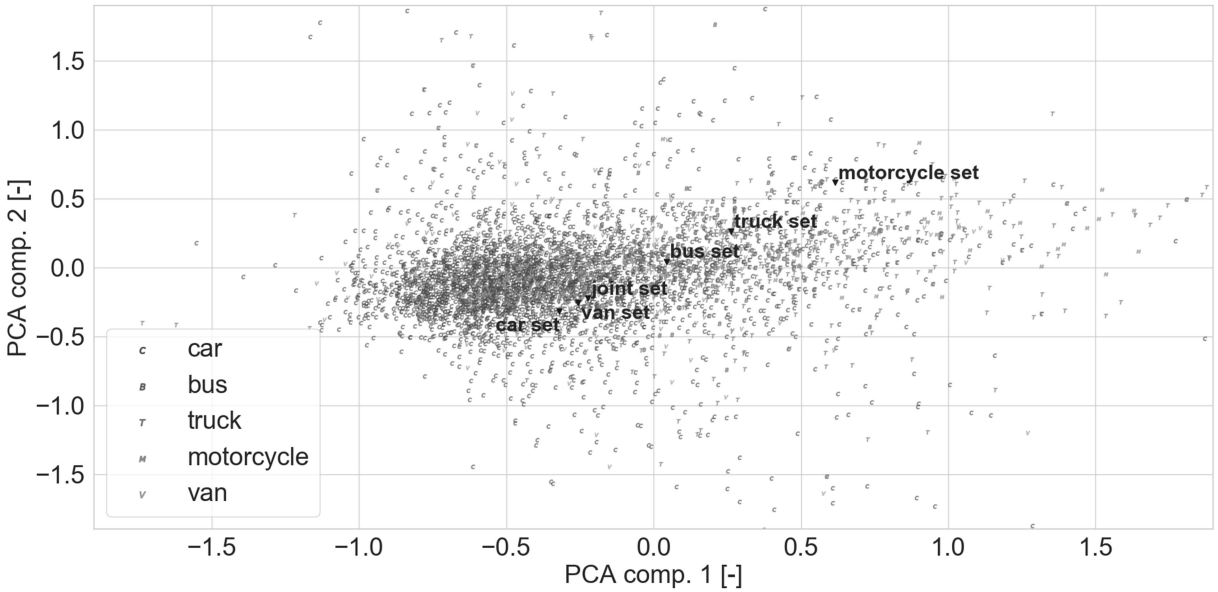

Figure 6. Result of the unsupervised training – decision space visualized with the use of the PCA transformation. Black triangle markers denote centroids of class-related data subsets and centroid of the whole joint dataset containing all data points from all subsets.

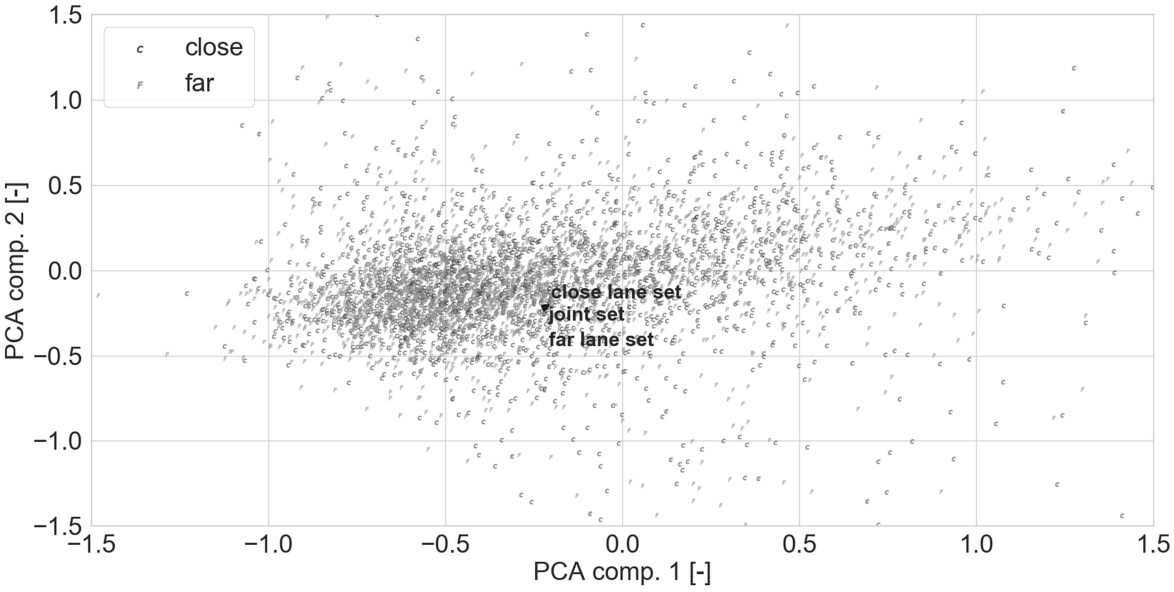

Figure 7. Result of the unsupervised training – decision space visualized with the use of the PCA transformation

A visual depiction of decision space shows, that there is possible to obtain separability for classes of a passing vehicle, however no similar clear tendency is visible for the type of lane the vehicle is placed on. To analyze if such separability occurs in the original high-dimensional decision space, a series of statistical tests was conducted. Such an approach is necessary, as PCA-based visualization show only a part of the original variance of the high-dimensional dataset. Data were split into seven subsets, each containing points associated with one of seven groups of vehicles. For each of those sets, a centroid was calculated in a similar manner as it is shown in Figures 6 and 7. Centroid was also calculated for the original set containing points from all the groups. In our statistical analysis, we used two distances – a distance to a centroid of a class-related cluster d_{class} and a distance to a centroid of a joint dataset d_{joint}. A visual depiction of the beforementioned distances used in this process is presented in Figure 8.

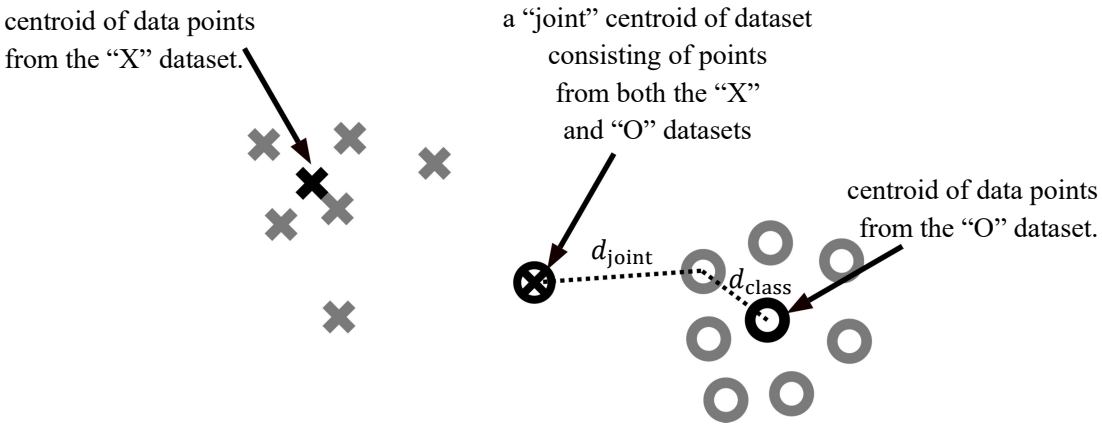

Figure 8. Depiction of calculation of distances from the centroid of the set containing examples of all classes (d_{joint}) and centroid of the set consisting only of examples of class to which belongs the currently analyzed point (d_{class}).=

Next, a difference between the distance to general set and to class-related set centroid was calculated, which can be written in short form as $\Delta d = d_{joint} - d_{class}$. The difference Δd is greater than zero if centroid of a class-related cluster of data points is positioned closer to the chosen data point than a centroid containing all analyzed points. Therefore, if statistically the mean or a median of Δd is greater than 0, then the separation between a joint dataset and a class-related dataset is obtained. This premise is used as an alternative hypothesis in statistical tests performed in our study. The null hypothesis is that the mean of Δd in a given dataset is zero and no separation is obtained. A boxplot of values of Δd derived in such a way is presented in Figure 9. The distance is normalized by the division of distance by a standard deviation of the joint dataset.

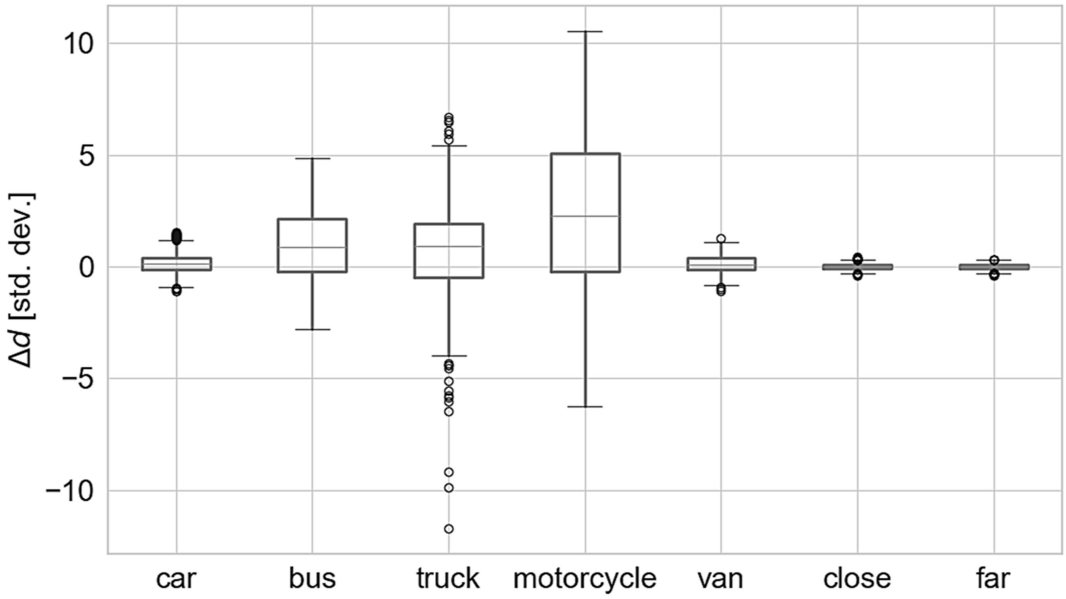

Figure 9. The difference of distances from the center of the set containing data points of all classes and center of the cluster containing only one class of data-points. The distance is expressed in terms of the mean standard deviation of the set containing all classes of vehicles.

To test statistical hypotheses about the obtained separability in the dataset generated by the autoencoder neural network, an iterated t-test was conducted to check if the mean value of distances depicted in Figure 9 is greater than zero. All p-values were lesser than a standard value of significance level $\alpha = 0.05$. Values of t-statistic, associated p-values and medians of Δd are given in Table 2. A Holm-Bonferroni correction for the multiple comparisons was applied to take into account the fact that iterated testing was performed.

Table 2. Values of t statistic obtained from the t-test (after application of Holm-Bonferroni multiple comparison corrections).

class	**car**	**bus**	**truck**	**motorcycle**	**van**	**close**	**far**
t statistic	23.45	6.90	7.35	6.67	7.32	2.32	1.97
p-value	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	0.041	0.049
Δd median [std. dev.]	0.149	0.874	0.904	2.300	0.111	0.001	0.002

As can be seen from Figure 9, the parameter which is unique to the motorcycle class is its wide interquartile range what suggests that data for this class are associated with high levels of variance when compared to the rest of classes. Therefore, a Brown-Forsythe statistic test was applied to test such a hypothesis. A resulting value of the Brown-Forsythe test statistic is 1543.2. Therefore, the p-value of the test is lesser than 10^{-3}, so we can conclude that there are statistically significant differences in the variance of distance differences. To test which pairs are statistically different in terms of variance a Dunn posthoc statistic test was performed. To test the variance of distances, the following transform was applied to the data:

$$\bar{d} = |d - E(d)| \tag{1}$$

The test was performed on the values of $\bar{d}$. Dunn's test performed on data transformed in such a way allows to find out a difference in the variance of two given datasets. Resulting p-values of the Dunn's test are provided in Table 3.

Table 3. Values of Dunn's posthoc test for equality of variances of two classes. Pairs for which a statistically significant ($\alpha = 0.05$) differences of variance found are marked with a bold font.

	car	bus	truck	motorcycle	van	close	far
car		**0.002**	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	0.540	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$
bus	0.002		0.869	0.433	**0.002**	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$
truck	$\mathbf{< 10^{-3}}$	0.869		0.393	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$
motorcycle	$\mathbf{< 10^{-3}}$	0.433	0.393		$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$
van	0.540	**0.002**	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$		$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$
close	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$		0.996
far	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	$\mathbf{< 10^{-3}}$	0.996	

It can be found that despite the fact that the labeled dataset contains only 93 examples, they represent a high variance and can be associated with high values of Δd. Therefore, there is a potential for obtaining a separation for this class. On the other hand, high variance means also that some examples may be hard to distinguish from other classes as an inter-quartile range of motorcycle-related distances overlaps with ranges associated with other analyzed classes.

The result of a supervised learning-based classification is shown in Table 4. As can be seen, in the table there are 4 classes of 7 discussed before. Since precision and recall scores for 3 remaining classes were close to zero, we decided to not include results for them, because such a result indicates poor accuracy. However, in case of detection of cars and trucks, and a vehicle presence on both: a lane which is closer or a lane more distant to the position of a camera, reasonable results of classification were achieved. An autoencoder output allowed training the neural network on both: labeled and unlabelled data.

There are some limitations of the presented approach. Firstly used dataset is unbalanced as there are differences of a number of examples provided for each class in the labeled dataset. This could cause the problem with the classification of poorly resented classes e.g motorcycle or bus. That can be seen in Table 4, especially if analyzing the F1 score results. This is probably the reason why no satisfactory result was obtained for the motorcycle class.

Table 4. Results of classification using the architecture presented above

	Accuracy	Precision	Recall	F1 score
car	0.870	0.706	0.703	0.704
truck	0.987	0.136	0.375	0.200
close lane	0.852	0.357	0.405	0.380
far lane	0.852	0.587	0.459	0.515

Secondly annotated dataset consists of only two hours of recordings, which could be not enough to properly capture all dependencies and features necessary for appropriate classification for use of autoencoder. The annotation process for traffic noise recordings is a time-consuming process, thus it does not enable acquiring more data within a reasonable time.

4. CONCLUSION

According to the presented results, with our approach, it is possible to classify cars with ~ 87% accuracy, with a satisfying precision and recall, as well with a ~70% F1 score. The remaining results are slightly worse, probably since numbers of examples in each class were unbalanced and most of the data were acquired for cars. Despite the above difficulties, a distinction between two classes of vehicles, namely: cars and trucks, was possible to make based on acoustical data. The outcome for motorcycle class may be a consequence of a low

number of observations for this class and the fact that any single-track vehicle was interpreted as a motorcycle. There is a lot of variety of single-track vehicles.

The obtained results of distinguishing between lanes may be influenced by the fact that the simultaneous occupation of both lanes occurred commonly in the analyzed data set. Based on the audio modality only, it is difficult to distinguish which lane is currently occupied, because the sounds from both lanes overlap, despite the differences. The solution to this problem could be the introduction of an additional class indicating the occupancy of both lanes in future work.

The advantage of the proposed approach is the fact, that the use of autoencoder for unsupervised pre-training allowed to utilize unlabelled audio data. Another advantage of the proposed classifier architecture lays in its modularity. The autoencoder-derived part of the network may be reused as a feature extraction module for another neural network performing the classification of vehicles. The frames of raw audio signals acquired in the proximity of a road can be used for both: supervised and unsupervised training of neural networks. However, despite an unsupervised pre-training stage it is still necessary to provide a sufficiently high number of annotated examples as it could be seen in the case of motorcycle class. Despite good characteristics in the decision space, the classifier was not able to achieve an entirely satisfactory accuracy level.

ACKNOWLEDGMENTS

The project entitled: "INZNAK: Intelligent Road Signs with V2X Interface for Adaptive Traffic Controlling (No. POIR.04.01.04-00-0089/16) is subsidized from the European Regional Development Fund by the Polish National Centre for Research and Development (NCBR)

REFERENCES

[1] X. V. Gonzalez and F. Alías, “Automatic classification of road vehicles considering their pass-by acoustic signature,” *Proc. Meet. Acoust.*, vol. 19, 2013.

[2] K. Marciniuk, B. Kostek, and A. Czyżewski, “Traffic Noise Analysis Applied to Automatic Vehicle Counting and Classification,” 2017, pp. 110–123.

[3] M. A. Sobreira-Seoane, A. Rodríguez Molares, and J. L. Alba Castro, “Automatic classification of traffic noise,” *Proc. - Eur. Conf. Noise Control*, no. June, pp. 6221–6226, 2008.

[4] M. A. D, B. S. A, and M. M. H, “Classification of Vehicles Based on Audio Signals using Quadratic Discriminant Analysis and High Energy Feature Vectors,” *Int. J. Soft Comput.*, vol. 6, no. 1, pp. 53–64, 2015.

[5] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, “auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks,” *J. Mach. Learn. Res.*, vol. 18, 2017.

[6] J. Chorowski, R. Weiss, S. Bengio, and A. Oord, “Unsupervised Speech Representation Learning Using WaveNet Autoencoders,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. PP, p. 1, 2019.

[7] F. Li *et al.*, “Feature extraction and classification of heart sound using 1D convolutional neural networks,” *EURASIP J. Adv. Signal Process.*, vol. 2019, p. 59, 2019.

[8] J. Kotus and G. Szwoch, “Calibration of acoustic vector sensor based on MEMS microphones for DOA estimation,” *Appl. Acoust.*, vol. 141, no. July, pp. 307–321, 2018.

[9] A. Kurowski, K. Marciniuk, and B. Kostek, “Separability Assessment of Selected Types of Vehicle-Associated Noise,” in *Multimedia and Network Information Systems*, 2017, pp. 113–121.

[10] K. Marciniuk, M. Szczodrak, and A. Czyzewski, “An application of acoustic sensors for the monitoring of road traffic,” *2018 Signal Process. Algorithms, Archit. Arrange. Appl.*, pp. 208–212, 2018.

[11] J. Kotus, “Determination of the Vehicles Speed Using Acoustic Vector Sensor,” in *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2018, pp. 64–69.

[12] A. Kurowski, A. Czyżewski, and S. Zaporowski, “SPA 2019 Automatic labeling of traffic sound recordings using autoencoder-derived features,” pp. 38–43, 2019.

[13] A. Czyżewski, “Free-standing intelligent road sign,” 125160, 2019 Polish patent No. W.125160,.

[14] A. Czyzewski, “Hanging intelligent road sign,” 125159, 2019. Polish patent No. W.125159

[15] N. Buduma and N. Locascio, *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*, 1st ed. O’Reilly Media, Inc., 2017.