Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 55 (2021) 1289–1295

Transportation Research Procedia

www.elsevier.com/locate/procedia

14th International scientific conference on sustainable, modern and safe transport

Detection of vowel segments in noise with ImageNet neural network architectures

René Fabricius[a,*] , Ondrej Šuch[a,b]

[a]*Žilinská Univerzita v Žiline, Univerzitná 8215/1, 010 26 Žilina, Slovakia*
[b]*Matematický ústav Slovenskej akadémie vied, Ďumbierska 1, 974 11 Banská Bystrica, Slovakia*

Abstract

In this article we report on experiments on detection of vowel segments in speech with additive noise. Deep neural networks have become the key algorithm in the majority of modern machine learning solutions. We investigate the performance of four ImageNet convolutional neural network (CNN) architectures. Usage of image processing CNNs is enabled by transforming the speech segments into spectrograms before the classification takes place. We perform experiments on TIMIT speech dataset and noise from datasets MAVD and ESC-50. The accuracy of individual architectures did not vary significantly among architectures on the dataset with added noise. However, accuracy of various architectures did differ significantly when applied to noise with absent speech.

Peer-review under responsibility of the scientific committee of the TRANSCOM 2021: 14th International scientific conference on sustainable, modern and safe transport

Keywords: Imagenet; deep neural network; vowel segments; additive noise; car audio systems

1. Introduction

Modern cars are increasingly equipped with sophisticated electronics including navigation and communication systems. An important research aim is to find the best way for a driver to interact with such systems. For many years, the primary interface to control computers has been via typing and mouse controls. Manual mode of control is not desirable in a car, because the driver's hands should stay on the steering wheel even if the car is in an (semi)-autonomous regime. Therefore, it would be very useful, if the driver could communicate his commands to the car's systems via voice interface provided by an automated speech recognition system (ASR system).

* Corresponding author.
E-mail address: rene.fabricius@fri.uniza.sk

Peer-review under responsibility of the scientific committee of the TRANSCOM 2021: 14th International scientific conference on sustainable, modern and safe transport
10.1016/j.trpro.2021.07.112

ASR systems have made tremendous progress in the past 50 years. An important early development was the introduction of hidden Markov models (HMM) for modelling of speech (Baker, 1975; Juang and Rabiner, 1991). More recently, deep neural networks transformed development of ASR systems by leapfrogging HMM systems (Hinton et al., 2012). Deep neural networks are nowadays a standard machine learning tool for both acoustic modelling as well as for language modelling. One major remaining challenge for ASR systems is recognition of speech in noisy environments. This problem is especially acute for ASR in transportation systems, since there are maybe many loud noise sources present such as engine noise, vibration, rain, wind, passenger crosstalk, etc.

Humans can recognize speech in extreme noise. Part of it can certainly be attributed to human understanding of the meaning of speech. However, linguistic experiments have shown that humans perform surprisingly well even when recognizing nonsensical syllables in noise, when a better language model cannot help (Parikh and Loizou, 2006). It is thus natural to ask, if the state-of-the-art deep neural networks could replicate human recognition performance in noisy environments.

A 2018 study investigated how the performance of convolutional neural networks degrades with noise for visual recognition tasks. The authors found that although the performance declines, it declines by small amounts and networks are partially resistant to a mismatch between noise levels during training and during testing phases (Hrabovsky et al., 2018).

In our study we carry out an analogous evaluation on the effect of noise on the performance of deep convolutional neural networks for processing of speech. We opted to focus on the problem of *vowel segment detection*. There are two reasons why we focus on this particular distinctive feature. First, these segments are among the most salient portions of the speech since air flows unobstructed through the supralaryngeal structures and retains most of its energy. Therefore, detection of these regions could be very advantageous for applications in ASR systems in noisy environments. Secondly, these segments are usually easy to visually identify in spectrograms due to the signature voiced indication and high energy. Therefore, one may expect that convolutional neural networks, the state-of-the-art approach in computer vision, should be suitable for their detection via transfer learning. We also note that detection of vowels has various other applications: for voice activity detection (Yoo and Yook, 2009), speech analysis and recognition (Prasanna et al., 2002), speaker identification (Daqrouq and Tutunji, 2014), emotion classification (Deb and Dandapat, 2017) and others.

Detection of vowels falls into a more general task of phonological features or articulatory features detection. Research in this field suggests that phonological features detection may be a preferable approach to phoneme detection in ASR as it is language independent and therefore generalizes better (Cerňak et al., 2015; Karaulov and Tkanov, 2019; King and Taylor, 2000). Use of neural networks in this field is not a novel idea. Applications of recurrent neural networks for phonological features detection were examined by (King and Taylor, 2000). (Cerňak et al., 2015) used a bank of multilayer perceptrons in an application of phonological vocoder. And in a more recent study (Karaulov and Tkanov, 2019) employed attention-based models for both articulatory features classification and phonemes classification.

2. Objectives

We decided to examine the performance of convolutional neural networks (CNN) for vowel detection in environments with additive noise (as opposed to noise created by reverberations). Our main goal was to evaluate several ImageNet architectures and to examine how each of them performs in this transfer learning task. There are several reasons why ImageNet architectures are particularly suitable for this task:

- ImageNet is one of the most studied problems in computer vision, and major innovations in deep network architectures are customarily evaluated on ImageNet dataset,
- ImageNet architectures are readily available in all major deep learning software (PyTorch, Keras, Matlab),
- ImageNet architectures have been used in transfer learning before,
- detection of vowel segments can make use of a relatively large spectrotemporal field, which fits well with processing ability of ImageNet networks that are trained for images of sizes exceeding 200 pixels.

In order to achieve the main goal, we also needed to introduce a benchmark task for noisy speech recognition which requires choosing additive noise.

3. Method

Our method combines speech analysis and image recognition techniques. Analyzed signal is first transformed into a spectrogram and this spectrogram is then fed into a convolutional neural network classifier. Output of the classifier is a binary variable conveying the information whether the center of the spectrogram belongs to a vowel segment.

Dataset preparation begins with construction of spectrograms. Each spectrogram is constructed from the spectra of 501 windows of length 32 ms sampled from the processed signal. These windows are overlapping and are sampled with a step of 2 samples which corresponds to 0.125 ms in a signal with a sampling rate of 16 kHz. The whole spectrogram therefore covers 94.5 ms of signal. To reduce spectral leakage, a window function is applied to every window before discrete Fourier transform takes place (Harris, 1978). The choice of the window function may affect the effectiveness of the whole method. Digital color images usually consist of three channels, that is three two-dimensional matrices, each corresponding to one primary color of light. Considering that a spectrogram is a single two-dimensional matrix of data, it is clear that we can create a color image from three separate spectrograms. This gives us a chance to exploit advantages of three different window functions and construct a more robust method. As the first window function, we have chosen the Gaussian window which is recommended by (Boersma, 1993) for a related task of fundamental frequency analysis. The remaining two window functions are Blackman-Harris window and Blackman-Harris-Nuttall window, both frequently used in spectrum analysis. Code used for dataset preparation can be found in a GitHub repository[1].

Detection part of the algorithm consists of the use of a convolutional neural network (CNN) (Lecun et al., 1998) for classification of the spectrogram. CNN are neural networks specifically designed for image processing. They are achieving state-of-the-art results in multiple computer vision tasks including image classification. Probably the most important benchmark for image classification is ImageNet dataset (Russakovsky et al., 2014). It consists of millions of images belonging to one thousand different classes. This dataset enabled creation of a multitude of various CNN architectures from which we selected four to be tested for the use in our method. The selected architectures in order of their publication are: AlexNet - (Krizhevsky, 2014), VGG - (Simonyan and Zisserman, 2014), ResNet - (He et al., 2016) and DenseNet - (Huang et al., 2017). The output of these CNNs are two values representing the predicted probabilities that the classified spectrogram was created from a vowel or not. Before training of the CNNs, we estimated a reasonable starting learning rate for each architecture by a method proposed in (Smith, 2017).This method suggests performing a training run with a low starting learning rate and to gradually increase it while observing both the learning rate and the training loss. The run can be stopped once training loss starts to increase rapidly. The reasonable starting learning rate is then suggested to be 1/10 of the learning rate for which the training loss function was at its lowest.

4. Experiments

We evaluate our method on speech data with added noise. For speech data we use the TIMIT dataset (Garofolo et al., 1992). TIMIT contains recordings of 630 speakers from eight dialects of American English. For noise data we choose the MAVD dataset (Zinemanas et al., 2019), which contains traffic noises from a busy urban environment and should therefore fit the transportation setting well.

CNNs which we use all perform a validation step in their training. Since TIMIT dataset only consists of train and test sets, we separate the train set into train and validation sets with dialects DR1-DR7 staying in the train set and dialect DR8 forming the validation set.

TIMIT dataset contains annotations for all phonemes in its speech data, which enables us to randomly sample points from either the vowel or the non-vowel portion of the speech. We center the windows for spectrogram creation

[1] https://github.com/ReneFabricius/VowelDetection

according to these points. Noise data are sampled also randomly. Speech and noise data are added together maintaining the desired signal to noise ratio (SNR). Vowel and non-vowel data have equal proportions in train, validation and test sets. Train and validation sets are generated with SNR uniformly distributed in the [-5 dB, 5 dB) interval. For testing we create five different sets, each with constant SNR of -6 dB, -3 dB, 0 dB, 3 dB and 6 dB.

Identical experiments are performed with each of the used CNN architectures. Selected CNN architectures and their parameters are displayed in Table 1.

Table 1. Parameters of used CNN architectures.

Architecture	Number of layers	Number of parameters	Starting learning rate
AlexNet	8	60M	0.005
VGG11	11	133M	0.0005
ResNet-18	18	11.5M	0.01
DenseNet-121	121	8.5M	0.1

We test each CNN on each of the five testing sets with different SNR. Results of the tests are displayed in Fig. 1. (left). As can be expected, the accuracy drops with decreasing SNR. Interesting observation is that the network ResNet-18, which has the best accuracy at SNR 6 dB, drops with the decreasing SNR to the last place. On the other hand, performance of the network VGG11, which is at second to the last place at SNR 6 dB, deteriorates with decreasing SNR the least and at SNR values of 3 dB and lower is the most accurate among the tested networks.

To test the robustness of our method against different noise than the training data were corrupted with we create another set of testing data. In this case we use noise data from the ESC-50 dataset (Piczak, 2015). This dataset contains 2000 recordings of various environmental sounds classified into 50 classes. Same as with MAVD dataset, we sample the noises randomly and create five test sets with SNR of -6 dB, -3 dB, 0 dB, 3 dB and 6 dB. Results of tests on these test sets are displayed in Fig. 1. (right). Tests are performed with the same instances of networks as the previous MAVD noise tests. Again, we can observe best robustness against decreasing SNR in network VGG11 and somewhat worse robustness in ResNet-18, which again is performing best among the tested networks at SNR 6 dB. Accuracy of networks VGG11 and ResNet-18 at SNR 6 dB and -6 dB for both types of noise is displayed in Table 2. As can be seen in the Table 2., accuracy at 6 dB dropped by more than 4% in both networks for ESC-50 noise as compared to MAVD noise with which the networks are trained. On the other hand, at SNR -6 dB accuracy of ResNet-18 slightly increased and accuracy of VGG11 dropped by 2.35% for ESC-50 noise as compared to MAVD noise. In conclusion it seems that robustness against unknown noise is for these networks better at lower SNR.

Table 2. Accuracies of two best performing networks.

Network	Noise dataset	Accuracy at SNR -6 dB (%)	Accuracy at SNR 6 dB (%)	Training accuracy (%)	Validation accuracy (%)
ResNet-18	MAVD	73.31	86.44	84.47	82.61
	ESC-50	73.40	82.28		
VGG11	MAVD	76.46	85.79	86.22	83.69
	ESC-50	74.11	80.84		

We performed a set of experiments on pure noise from the used noise datasets to see how often erroneous detection of a vowel in noise occurs. Experiments are done by uniformly sampling approximately 100,000 points from both noise datasets and by classifying spectrograms centered at these points. Results of these experiments are displayed in Fig. 2. Different architectures perform differently at this task. ResNet-18 has the lowest error rate by a large margin for both noise datasets. We can also see that erroneous detection in the MAVD dataset occurs less frequently than in the ESC-50 dataset. Average misclassification rate over all networks in MAVD dataset is 6.92% whereas in the ESC-50 dataset it is 30.78%. This result can be mainly attributed to the presence of sounds like church bells, siren, sheep or crying baby in the ESC-50 dataset.

To get a visual representation of the results produced by our method, we test it on a whole sentence from TIMIT dataset. We pick a sentence which does not contain any voiced parts except for vowels to be able to compare our

results with Praat software[2] pitch analyzer. We selected the sentence “Perfect, he thought.”, in the TIMIT dataset it can be found in the file TRAIN/DR8/MTCS0/SI1972.WAV. We added to it a MAVD dataset noise from the beginning of the file audio_test/templete0103_02.flac at SNR of 6 dB. We created spectrograms from the resulting sound with a step of 10 ms and then classified each of them using the ResNet-18 network. Result of this experiment is displayed in Fig. 3. Our classifications are corresponding with the TIMIT annotations well, except for a small segment of vowel detected in the interval [0.60125 s, 0.60625 s] where there is a pause between words “Perfect” and “he”.

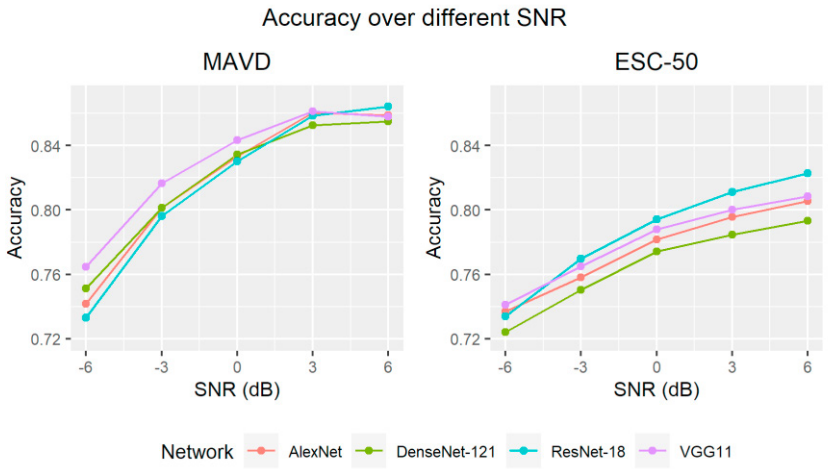

Fig. 1. Accuracy of tested CNNs over different SNR with MAVD noise (left) and ESC-50 noise (right).

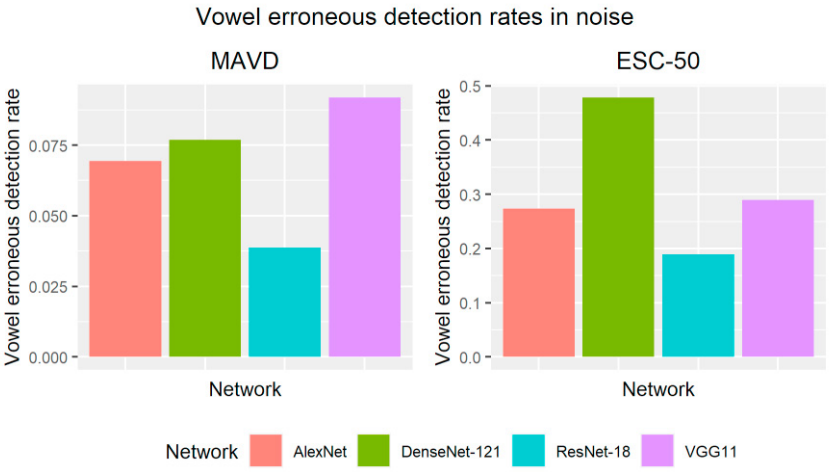

Fig. 2. Analysis of the noise data, MAVD noise on the left and ESC-50 noise on the right. Bars represent erroneous detection rates of vowels in clear noise data for tested CNNs.

[2] https://github.com/praat/praat

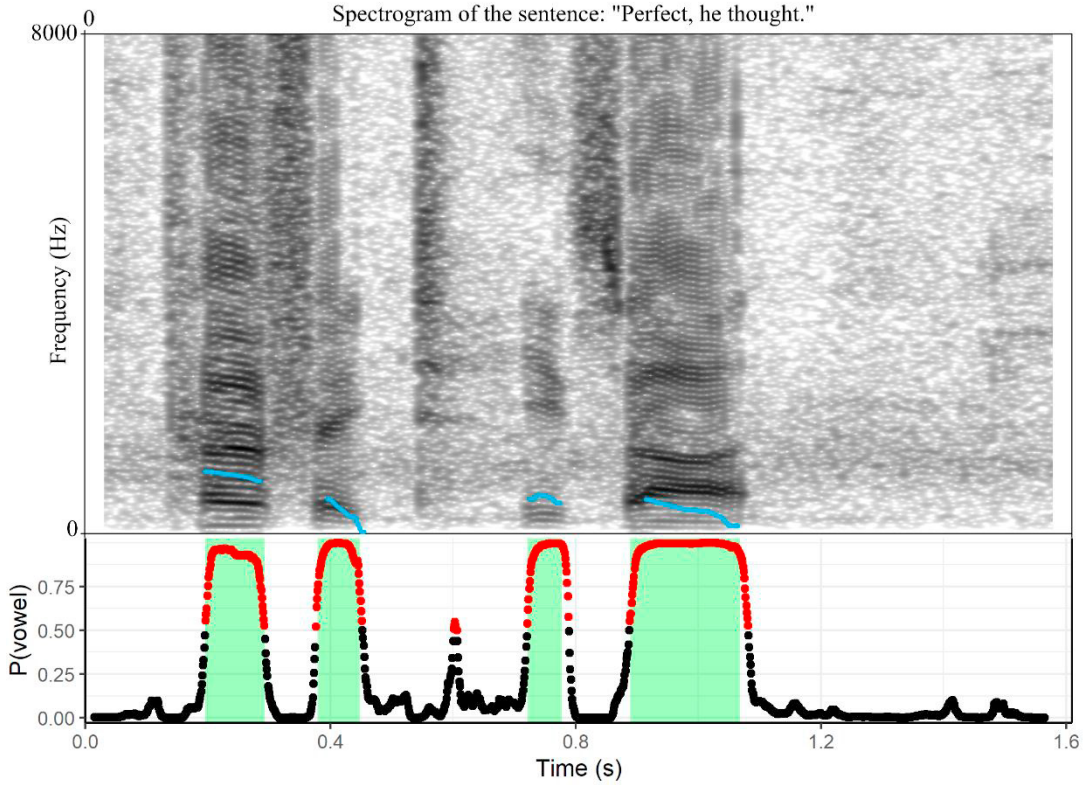

Fig. 3. Comparison of our method with Praat pitch analysis. In the top part of the image is the spectrogram generated by Praat with detected pitch displayed as blue lines. Results of our method are displayed in the bottom part. Each point represents a performed classification. Horizontal coordinate of the point gives the time in which the classified spectrogram was centered and the vertical coordinate gives the probabilistic result of the classification. Points with classification output of being a vowel higher than 50% are displayed in red. Background of the plot is filled with green in the intervals where vowel phonemes are present according to TIMIT annotations.

5. Conclusion

ImageNet neural network architectures were able to detect vowel segments in noise with accuracy ranging between 76-86% depending on the SNR which ranged from -6dB to 6dB. For comparison, a recent study with attention networks achieved 70% accuracy for vowel detection for speech without any noise (Karaulov and Tkanov, 2019). The accuracy of individual architectures did not vary significantly among architectures on the dataset with added noise.

However, accuracy of various architectures did differ significantly when applied to noise with absent speech. The best performing was ResNet-18 network architecture which erroneously classified almost 20% of noise as vowels. This is still a somewhat high error rate and indicates that the task of detection of vowels in noise is nontrivial and may require design of bespoke network architectures or possibly significantly more training data.

Acknowledgements

We would like to thank J. Juhár for suggestions to improve our paper. Our research was partially supported by VEGA grant 2/0144/18.

References

Baker, J., 1975. The DRAGON system–An overview. IEEE Transactions on Acoustics, Speech, and Signal Processing 23, 24–29. https://doi.org/10.1109/TASSP.1975.1162650

Boersma, P., 1993. ACCURATE SHORT-TERM ANALYSIS OF THE FUNDAMENTAL FREQUENCY AND THE HARMONICS-TO-NOISE RATIO OF A SAMPLED SOUND.

Cerňak, M., Potard, B., Garner, P., 2015. Phonological vocoding using artificial neural networks. pp. 4844–4848. https://doi.org/10.1109/ICASSP.2015.7178891

Daqrouq, K., Tutunji, T., 2014. Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers. Applied Soft Computing 27. https://doi.org/10.1016/j.asoc.2014.11.016

Deb, S., Dandapat, S., 2017. Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions. IEEE Transactions on Affective Computing PP, 1. https://doi.org/10.1109/TAFFC.2017.2730187

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V., 1992. TIMIT Acoustic-phonetic Continuous Speech Corpus. Linguistic Data Consortium.

Harris, F.J., 1978. On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE 66, 51–83. https://doi.org/10.1109/PROC.1978.10837

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Hinton, G., Deng, li, Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. Signal Processing Magazine, IEEE 29, 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hrabovsky, J., Kontsek, M., Segeč, P., Such, O., 2018. Influence of Positive Additive Noise on Classification Performance of Convolutional Neural Networks. pp. 175–180. https://doi.org/10.1109/DISA.2018.8490611

Huang, G., Liu, Z., Maaten, L. van der, Weinberger, K.Q., 2017. Densely Connected Convolutional Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Juang, B.H., Rabiner, L.R., 1991. Hidden Markov Models for Speech Recognition. Technometrics 33, 251–272. https://doi.org/10.1080/00401706.1991.10484833

Karaulov, I., Tkanov, D., 2019. Attention model for articulatory features detection.

King, S., Taylor, P., 2000. Detection of Phonological Features in Continuous Speech using Neural Networks. Computer Speech & Language 14, 333–353. https://doi.org/10.1006/csla.2000.0148

Krizhevsky, A., 2014. One weird trick for parallelizing convolutional neural networks.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791

Parikh, G., Loizou, P., 2006. The influence of noise on vowel and consonant cues. The Journal of the Acoustical Society of America 118, 3874–3888. https://doi.org/10.1121/1.2118407

Piczak, K.J., 2015. ESC: Dataset for Environmental Sound Classification. https://doi.org/10.7910/DVN/YDEPUT

Prasanna, S., Gangashetty, S., Yegnanarayana, B., 2002. Significance Of Vowel Onset Point For Speech Analysis.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L., 2014. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115. https://doi.org/10.1007/s11263-015-0816-y

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

Smith, L.N., 2017. Cyclical Learning Rates for Training Neural Networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 464–472. https://doi.org/10.1109/WACV.2017.58

Yoo, I.-C., Yook, D., 2009. Robust Voice Activity Detection Using the Spectral Peaks of Vowel Sounds. Etri Journal - ETRI J 31, 451–453. https://doi.org/10.4218/etrij.09.0209.0104

Zinemanas, P., Cancela, P., Rocamora, M., 2019. MAVD: A Dataset for Sound Event Detection in Urban Environments. pp. 263–267. https://doi.org/10.33682/kfmf-zv94