Sounding Industry: Challenges and Datasets for Industrial Sound Analysis (ISA)

Sascha Grollmisch
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany
goh@idmt.fraunhofer.de

Jakob Abeßer
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
abr@idmt.fraunhofer.de

Judith Liebetrau
Industrial Media Applications
Fraunhofer IDMT
Ilmenau, Germany
ltu@idmt.fraunhofer.de

Hanna Lukashevich
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
lkh@idmt.fraunhofer.de

***Abstract*—The ongoing process of automation in production lines increases the requirements for robust and reliable quality control. Acoustic quality control can play a major part in advanced quality control systems since several types of faults such as changes in machine conditions can be heard by experienced machine operators but can hardly be detected otherwise. To this day, acoustic detection systems using airborne sounds struggle due to the highly complex noise scenarios inside factories. Machine learning systems are theoretically able to cope with these conditions. However, recent advancements in the field of Industrial Sound Analysis (ISA) are sparse compared to related research fields like Music Information Retrieval (MIR) or Acoustic Event Detection (AED). One main reason is the lack of freely available datasets since most of the data is very sensitive for companies. Therefore, three novel datasets for Industrial Sound Analysis with different application fields were recorded and published along with this paper: detection of the operational state of an electric engine, detection of the surface of rolling metal balls, and detection of different bulk materials. For each dataset, neural network based baseline systems were evaluated. The results show that such systems obtain high classification accuracies over all datasets in some of the subtasks which demonstrates the feasibility of audio-based analysis of industrial analysis scenarios. However, the baseline systems remain highly sensitive to changes in the recording setup, which leaves a lot of room for improvement. The main goal of this paper is to stimulate further research in the field of ISA.**

***Index Terms*—industrial sound analysis, machine learning, audio, signal processing, deep learning, neural networks, datasets**

I. Introduction

With the ongoing process of digitalization in the industry, many processing lines are getting automated. For end-of-line testing permanent quality control mechanisms are required to ensure that no faulty products are produced and shipped. Predictive maintenance using Industrial Sound Analysis (ISA) aims at detecting subtle changes in airborne sound that indicate potential machine faults early enough to avoid expensive downtime in production lines. Commonly applied fault detection systems are based on cameras, structure-borne sensors, or monitoring of production line parameters such as current and temperature. But in fact, faults such as broken ball bearings or defective engines can be easily heard and recognized by experienced machine operators but hardly (camera) or only costly (X-ray, ultrasound) recognized otherwise. However, detecting audible errors in the human hearing range is very challenging due to highly complex noise scenarios in fabrication plants. Therefore, the main challenge of ISA is to be robust against unpredictable background noise in factories or other industrial scenarios while being sensitive to subtle differences in the observed sound.

II. Related Work

Two related research fields to ISA are machine listening and Music Information Retrieval (MIR). In machine listening, audio scene classification and Audio Event Detection (AED) deals with classifying environmental sounds with a high variance, e. g., bus and tram, in real world scenarios like public places. In the last years, AED algorithms improved notably due to bigger datasets and the recent advances in machine learning and deep learning in particular [1]–[4]. In MIR, many research tasks deal with detecting, separating, and classifying individual components such as musical notes from complex mixtures of multiple musical instruments playing simultaneously. As shown for instance in the annual algorithm competition Music Information Retrieval Evaluation eXchange (MIREX) [5], novel deep learning architectures and a multitude of available audio datasets [6]–[8] helped the community to significantly improve results in many tasks.

ISA shares the complexity of the analyzed audio backgrounds with MIR and AED but the main task differs. In ISA, the sound differences that need to be detected are more subtle compared to AED signals. Also, the analyzed signals are commonly less harmonic and more noisy compared to MIR scenarios. Thus, not all datasets and algorithms can be easily adopted and applied for ISA which shows the necessity of dedicated datasets for achieving comparable advancements. One main reason for the lack of publicly available ISA datasets is the sensitivity of the recorded data. Many companies do not want the sounds of their products or additional metadata like rejection rates to be published. Not only the data itself but also the applied algorithms are often part of a company's intellectual property and are therefore kept secret. Only few publications on ISA can be found [9]–[14] and to best of our knowledge none of them published datasets. This lack of datasets hinders progress and is a strong disadvantage for

smaller companies not being able to afford comprehensive research.

III. Industrial Sound Datasets

Three novel datasets have been created and published along with this paper under CC BY 4.0 license[1] to stimulate broad research and cross testing of ISA algorithms. Each dataset represents an unique application scenario. All recordings are labeled with the selected classes in each task making them supervised classification problems [15]. All files were recorded with a sample rate of 44100 Hz and a resolution of 32bit and saved as wave files.

A. Electric Engine

1) Application Overview: Small electric engines are used in several applications like industrial fans, machine tools, car seats, household appliances, and different power tools while large electric engines can be found in pipeline compression and pumped-storage applications with ratings reaching 100 megawatt. For all areas of use it is important that the machines are running properly and the applied loads do not destroy them. Many failure types can neither be detected visually nor using structure-borne sound sensors, which are also often hard to apply especially to smaller engines. During runtime however, these faults become often audible.

This dataset is intended as a basis for developing algorithms which could later be integrated in production lines. Three similar units of an electric engine [2] were manipulated to simulate different acoustic conditions. The first engine is operated by 60 % of supply voltage and represents the operational state "good". For the second engine, the supply voltage changes every 18 ms between 15 % and 75 % of supply voltage to simulate a "broken" unit. An additional weight was applied to the third engine, also powered by 60 % of supply voltage, which represents the "heavy load" scenario. All engines were placed in separated plastic casings simulating the integration in a closed product and combined in a bigger case to suppress surrounding noise sources while testing. This approach could be implemented both in factories as well as in end-of-line-testing scenarios. The three engines were mounted close to each other in the case and can be switched on and off with a central switch placed outside of the case. As shown in Figure 1, the microphone [3] used for recording the engines was placed in a central position.

2) Dataset Overview: In March 2017 the `IDMT_ISA_ELECTRIC_ENGINE` dataset was recorded at the Fraunhofer Institute for Digital Media Technology (IDMT). In each file, only one of the engines is active at the same time assuming an engine can only have one of the three operational states discussed in the previous section. The dataset contains the following background noise types:

Fig. 1. Electric engine setup.

- `pure` - recordings with no additional background noise
- `talking` - people talking around the casing
- `white_noise` - white noise played back using speakers outside of the casings
- `atmo` - atmospheric sounds from a factory environment at three loudness levels (`low`, `medium`, `high`) played back using speakers
- `stress test` - with slightly changed input gains for simulating manipulations and people knocking on the casing

The training set only contains the pure recordings. The remainder of the recordings is intended as test set, which allows to test novel algorithms for their robustness against various background noise types and levels. The recordings are additionally provided in a pre-cut version as 3 second long segments for easier data shuffling and evaluation over shorter time periods.

B. Metal Ball Surface

1) Application Overview: Due to friction metal surfaces may suffer from abrasion or production problems may lead to already damaged surfaces. As an example scenario, we consider ball bearings where a polished surface of the included metal balls is important to ensure efficiency. Acoustic quality control can be applied since potential material damages are audible but the metal balls within the bearing cannot be accessed with other sensors. This dataset was created to detect the surface condition of an steel metal ball rolling down a slide made out of steel using a low-cost microphone [4] as sensor. As shown in Figure 2, a launch pad is attached to the slide, which releases up to three metal balls. Depending on the applied pressure, the metal balls roll over a coated steel slide[5] at slightly different speeds while passing the microphone. Due to the varying speed it may happen that multiple metal ball are on the slide simultaneously. A casing was placed around the slide in order to damp unwanted background noise. Three types of metal balls were produced with five units each: "Eloxed" (green), "Coated" (dark grey, using Diamor®[5]), and "Broken" (bronze, surface treated with sandpaper). The "Eloxed" balls

[1] creativecommons.org/licenses/by-sa/4.0/

[2] ACT Motor Brushless DC 42BLF01, 4000 RPM, 24VDC

[3] self built microphone based on an analog electret condenser mic capsule with following parameters: frequency range 50 Hz to 20 kHz, voltage range 2 V to 10 V, omni directional, sensitivity -35 dB ±4 dB

[4] MicW i456, www.mic-w.com/product.php?id=24

[5] coated with Diamor®by Fraunhofer IWS, see www.iws.fraunhofer.de/content/dam/iws/en/documents/publications/product_sheets/500-1_diamor_en.pdf

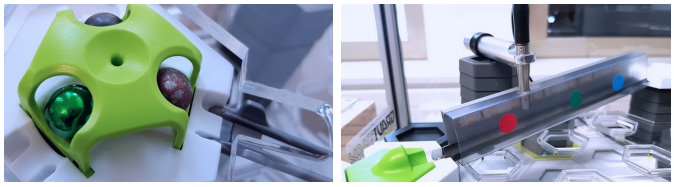

Fig. 2. Launch pad with all ball types (left). Slide with mounted microphone (right).

are manufactured on an industry scale and are the basis for the two other types. Since coating and breaking where applied manually to the original balls these types vary stronger. The audio recordings were cut based on the detected impact sound when a ball exits the slide. Files where multiple balls nearly touch each other were discarded due to overlapping sounds.

2) Dataset Overview: The `IDMT_ISA_METAL_BALLS` dataset was recorded at Fraunhofer IDMT in February 2018. In order to simulate a realistic acoustic environment, factory sound recordings were played back as background sounds using external loudspeakers during the recording process. The dataset was split into test and training set. We provide both an unbalanced split, which resulted from the original recording process, with a total size of 2267 recordings as well as a balanced split for reproducibility of experiments. Two additional test sets with more than 250 recordings each are provided where general settings like the angle of the slide and the position of microphone have been altered for testing the robustness of the system to changes in the setup.

C. Tubes

1) Application Overview: Different application scenarios with respect to filling of goods are imaginable. Firstly, the monitoring of unwanted altering of products during filling process. Secondly, the sorting of different materials and thirdly, the filling of an exactly defined amount of bulk material. We created this dataset in order to analyze to what extend it is possible to distinguish between different kinds of bulk material solely by using audio analysis. As shown on the left side of Figure 3, a printed slide was mounted into a plastic tube to which the same type of microphone as in section III-A has been mounted. The bulk material is inserted manually into a hopper on the tube which results in varying speeds of the bulk material rolling down the slide. This and the different amount of bulk material influences the sound and increases the complexity of the task. The sound, produced by the sliding of the material, is recorded. As second tube, identical in construction, was used for preparation of additional data in order to test the robustness of a recognition algorithm against differences in the manufacturing process. As shown on the right side of Figure 3, the bulk materials consist of five types of sweets and five kinds of nuts and screws.

2) Dataset Overview: The `IDMT_ISA_TUBES` dataset has been recorded and annotated March and April 2017 at Fraunhofer IDMT. Two tubes with identically built microphones were used for this dataset. Each bulk material has been poured

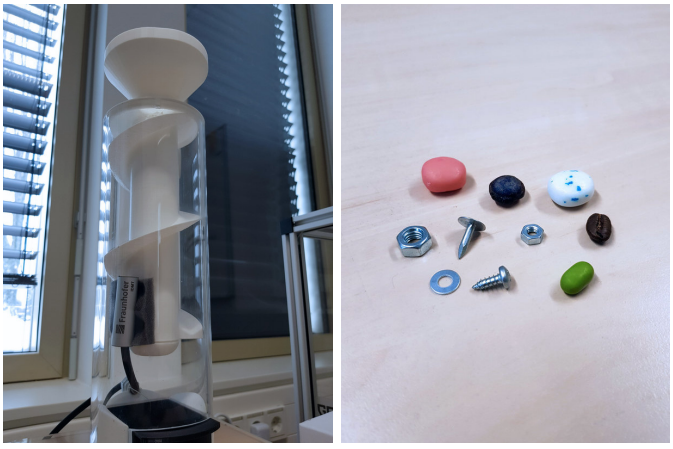

Fig. 3. Tube with slide and mounted external microphone (left). Bulk material (right).

50 to 60 times in each tube summing up to 1076 files in total. Each file has been cut by hand starting at the filling process until all items reached the end of the tube. Due to the small amount of examples per class for each tube no separate training and test sets are provided and cross-validation is recommended. The distribution that was used for each cross-validation step is provided with the dataset.

IV. BASELINE EXPERIMENTS

Deep Neural Networks (DNN) [16] were selected as baseline systems for all three use-cases as they a good ratio between classification accuracy and computational cost [17] which is important for industrial applications in terms of performance on embedded devices. By this, lowers latency can be achieved and security risks compared to more expensive cloud solutions are avoided. For the sake of simplicity, we compute magnitude frames from a FFT using a Hann window function and 50 % overlap as input feature representation for the DNN models. The phase is discarded here. The magnitude frames were normalized to a range of $[0, 1]$ based on the maximum value of the training data. This value was used for normalizing both the training and the test data. A random selection of 90 % of the training data was used for training the model while the remaining 10 % were used as validation data. Each experiment was repeated five times with a different random shuffling between training and validation set.

We provide both frame-wise accuracy A_{frame} and file-wise accuracy A_{file} as evaluation measure. For the latter, each file is classified based on a majority decision over all frame-wise classification results. The frame-wise accuracy shows how the system will perform in a real-time application scenario with the lowest latency possible. The file-wise accuracy on the other hand demonstrates how well the system performs in an offline scenario where a longer time-period is analyzed. This scenario is suitable for many end-of-line tests where the result of the whole measurement is more important than short-term variances in single time frames. The reported accuracy values

are the average over all accuracy values from the individual cross-validation folds after optimization.

For each task, model hyperparameters such as the number of layers, units per layer, dropout ratio [18], and FFT window size were optimized using Bayasing Optimization Techniques [19] [6]. All networks were trained for 1000 epochs using the Adam optimizer [20] with a learning rate of 0.001 and the categorical cross-entropy as cost function. As activation functions, we used ReLU [21] in all layers but the final one, where the softmax function is used. All weights are initialized using Glorot uniform initializer [22]. All experiments were conducted using the Keras framework with Tensorflow as backend[7].

The final DNN architecture for each task is reported in Table I. The "Input" column provides the number of input features to the model, which is the half of the FFT size. The "Architecture" column gives both the optimal number of layers and the number of units per layer. The best dropout ratio is provided between input and the first hidden layer ("DIn"), the remaining hidden layers ("D"), as well as before the final classification layer ("DOut").

TABLE I
OPTIMAL DNN ARCHITECTURES FOR EACH DATASET.

Dataset	Input	Architecture	DIn	D	DOut
Electric Engine	2048	256, 64, 16, 3	0.2	0.5	0.2
Metal Ball Surface	4096	1024, 256, 3	0.3	0.5	0.5
Tubes	4096	64, 10	0.0	0.5	0.0

A. Electric Engine

Table II summarizes the classification results obtained for the electric engine dataset with the corresponding model from Table I.

TABLE II
BASELINE RESULTS FOR IDMT_ISA_ELECTRIC_ENGINE.

Test set	A_{frame}	A_{file}
talking	95.1%	96.7%
white_noise	99.3%	100%
atmo_low	100%	100%
atmo_medium	99.9%	100%
atmo_high	94.7%	100%
stress test	82.3%	86.7%

The results demonstrate that DNNs and airborne sound can be used for classifying the operational state of an electric engine. Even though only being trained on pure recordings, the system performs fairly well on most noise types and unknown test conditions. However, it has to be considered that the engines sounds included in the dataset vary stronger than engines in real use-cases. As expected, increasing the noise level (atmo low to atmo high) decreases the classification performance. Changing the input gain and adding noise which leads to clipped signals (stress test) drops the accuracy under 90 % clearly leaving room for improvement.

[6]Implementation taken from github.com/fmfn/BayesianOptimization

[7]Keras: keras.io, Tensorflow: www.tensorflow.org

B. Metal Ball Surface

Table III summarizes the classification results obtained on the balanced dataset of the metal ball surface use case with the corresponding model from Table I.

TABLE III
BASELINE RESULTS FOR IDMT_ISA_METAL_BALLS.

Test set	A_{frame}	A_{file}
test	92.0%	98.8%
variation_set1	50.2%	51.4%
variation_set2	56.3%	59.2%

The results show that the surface of the metal ball can be predicted using airborne sound and neural networks with a high accuracy. Especially when looking at the full recording (A_{file}), the accuracy is close to 99%. However, slight variations of the setup have a huge impact on the classification performance. As shown in table IV, mainly the broken ball was misclassified.

TABLE IV
CONFUSION MATRIX FOR VARIATION_SET2.

True / Pred	varnished	coated	broken
varnished	88.0%	3.9%	8.1%
coated	16.5%	70.0%	13.5%
broken	47.7%	32.7%	19.6%

C. Tubes

Table V provides the classification results obtained on the tubes dataset with the corresponding model from Table I. One can observe a simple two-layer neural network achieves good accuracy scores up to 79% on a frame-level. For the Tube1 and Tube2 test sets, aggregating the frame-wise results over full files with an average length of 3s leads to perfect classification results. On a frame-level, there is still room for improvement. As can be seen in the lower two rows of Table I, such as simple model remains very sensitive to small changes in the setup as also observed for the other two use-cases.

TABLE V
BASELINE RESULTS FOR IDMT_ISA_TUBES.

Test set	A_{frame}	A_{file}
Tube1	79.1%	100.0%
Tube2	72.1%	100.0%
Tube1 train, tube2 test	34.1%	51.6%
Tube2 train, tube1 test	39.7%	48.3%

V. CONCLUSIONS

This paper defines closer the novel research of Industrial Sound Analysis (ISA) which deals with analyzing industrial airborne sounds for automated quality control, and sets it into

context with audio-related research fields such as Music Information Retrieval and Acoustic Event Detection. We discussed that nowadays, one main obstacle for advancing analysis algorithms in the field of ISA is the lack of publicly available datasets. Therefore, we introduced and published three novel datasets alongside with this paper.[8] The datasets cover three different application fields: detection the operational state of an electric engine, analyzing the surface of metal balls, and detection the type of bulk material. For each dataset, a neural network based baseline system has been evaluated. While these baseline systems achieved high classification accuracy under noisy conditions, especially for engines and metal balls, they reveal problems when the recording setup or observed objects change. Enhancing the robustness against these changes should be a focus on upcoming research. Possible advancements could benefit from data augmentation techniques or different input features and networks architectures better fitting ISA data, like convolutional or recurrent neural networks. Furthermore, the published datasets can be used for few-shot learning approaches by reducing the required number of training files. Also the task of semi-supervised classification can be targeted by using only "good" examples for training and testing the deviation to the other classes.

ACKNOWLEDGMENT

We would like to thank our colleagues from the *Branch Hearing, Speech and Audio Technology in Oldenburg* for providing us with the tubes and microphones used for the tubes and electric engine dataset. For the metal ball surface dataset, we thank our colleagues from Fraunhofer IWS for coating the metal balls and slides and Ravensburger AG for supplying us with the GraviTrax sets. Finally, we would like to thank our research assistants for supporting us in the sometimes tedious task of data acquisition.

REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Cham: Springer International Publishing, 2018.

[2] T. Heittola, A. Mesaros, and T. Virtanen, "TUT urban acoustic scenes 2018, development dataset," Apr 2018.

[3] J. Salamon, C. Jacoby, and J. P. Bello, "Urban Sound Datasets," in *Proc. 22nd ACM Int. Conf. Multimed.*, no. 3, 2014, pp. 1041–1044.

[4] Y. Choi, O. Atif, J. Lee, D. Park, and Y. Chung, "Noise-Robust Sound-Event Classification System with Texture Analysis," *Symmetry (Basel).*, vol. 10, no. 9, p. 402, sep 2018.

[5] J. S. Downie and Y. Hao, "MIREX 2017 Evaluation Results," School of Information Sciences, University of Illinois at Urbana-Champaign, Tech. Rep., 2017.

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of the 42nd International Conference on Acoustics, Speech, and Signal (ICASSP)*, New Orleans, LA, USA, 2017, pp. 776–780.

[7] A. Lerch, "Audio Content Analysis - Datasets," (last accessed 24.01.2019). [Online]. Available: http://www.audiocontentanalysis.org/data-sets/

[8] ISMIR, "ISMIR resources," (last accessed 24.01.2019). [Online]. Available: http://ismir.net/resources.html/

[8] Link will be placed here for the accepted camera-ready version

[9] P. Lipar, M. Cudina, P. Steblaj, and J. Prezelj, "Automatic Recognition of Machinery Noise in the Working Environment," *Strojniki Vestn. - J. Mech. Eng.*, vol. 61, no. 12, pp. 698–708, dec 2015.

[10] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised Detection of Anomalous Sound based on Deep Learning and the Neyman-Pearson Lemma," *IEEE/ACM Trans. Audio Speech Lang. Process.*, pp. 1–13, 2018.

[11] W. Li and C. K. Mechefske, "Detection of Induction Motor Faults: A Comparison of Stator Current, Vibration and Acoustic Methods," *J. Vib. Control*, vol. 12, no. 2, pp. 165–188, feb 2006.

[12] P. Henriquez, J. B. Alonso, M. A. Ferrer, and C. M. Travieso, "Review of Automatic Fault Diagnosis Systems Using Audio and Vibration Signals," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 44, no. 5, pp. 642–652, may 2014.

[13] E. Cano, J. Nowak, and S. Grollmisch, "Exploring Sound Source Separation for Acoustic Condition Monitoring in Industrial Scenarios," in *2017 24th Eur. Signal Process. Conf.*, Kos Island, Greece, 2017.

[14] F. Yang, M. S. Habibullah, T. Zhang, Z. Xu, P. Lim, and S. Nadarajan, "Health Index-Based Prognostics for Remaining Useful Life Predictions in Electrical Machines," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 4, pp. 2633–2644, apr 2016.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., ser. Information Science and Statistics. Springer-Verlag New York Inc, 2006.

[16] Y. Lecun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] S. Sigtia, A. M. Stark, S. Krstulovic, and M. D. Plumbley, "Automatic Environmental Sound Recognition: Performance versus Computational Cost," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2096–2107, jul 2016.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[19] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 2951–2959. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999325.2999464

[20] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6, dec 2014.

[21] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018.

[22] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*, 2010.