Using Multimodal Contextual Process Information for the Supervised Detection of Connector Lock Events

David Bricher(✉) and Andreas Müller

Institute of Robotics, Johannes Kepler University, Linz, Austria
david.bricher@bmw.com, a.mueller@jku.at

Abstract. The field of sound event detection is a growing sector which has mainly focused on the identification of sound classes from daily life situations. In most cases these sound detection models are trained on publicly available sound databases, up to now, however, they do not include acoustic data from manufacturing environments. Within manufacturing industries, acoustic data can be exploited in order to evaluate the correct execution of assembling processes. As an example, in this paper the correct plugging of connectors is analyzed on the basis of multimodal contextual process information. The latter are the connector's acoustic properties and visual information recorded in form of video files while executing connector locking processes.

For the first time optical microphones are used for the acquisition and analysis of connector sound data in order to differentiate connector locking sounds from each other respectively from background noise and sound events with similar acoustic properties. Therefore, different types of feature representations as well as neural network architectures are investigated for this specific task.

The results from the proposed analysis show, that multimodal approaches clearly outperform unimodal neural network architectures for the task of connector locking validation by reaching maximal accuracy levels close to 85%. Since in many cases there are no additional validation methods applied for the detection of correctly locked connectors in manufacturing industries, it is concluded that the proposed connector lock event detection framework is a significant improvement for the qualitative validation of plugging operations.

Keywords: Connector lock detection · Manufacturing sound events · Sound event detection · Applied machine learning · Neural networks · Optical microphone · Deep learning

1 Introduction

Although the degree of automation in manufacturing industries is constantly increasing, the correct plugging and locking of cable harness connectors is still a challenging task for machines [5,11] and is therefore mainly carried out manually.

I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 584, pp. 123–134, 2020.
https://doi.org/10.1007/978-3-030-49186-4_11

As the human failure rate for manual working tasks typically exceeds the failure rate of automated processes, additional validation methods are introduced, such that possible errors can be detected before leaving the manufacturing line.

In many cases the validation of correctly assembled parts can be achieved with visual inspection, but the investigation of correctly assembled connector cables solely based on image data is rather inefficient due to multiple reasons, e.g. occlusion by other assembling parts, variation of connector positions or drill of connector cables. Moreover, the validation of plugging processes is often not carried out at all.

Thus, it is of great interest to find multimodal contextual process information which can be used for the validation of correctly executed connector plugging processes. The most common errors occurring are those where the connectors have not been locked properly. A potential approach to assess, whether a locking has been correctly performed, is to analyze the inherent acoustic properties of the plugging event. Consequently, the presence and correct classification of locking sound events can provide information on the qualitative execution of plugging processes.

The field of sound event detection is a challenging sector, whose main target is to mimic the human capability of distinguishing different acoustic events and correctly classifying them. In the last decade multiple approaches (e.g. Gaussian mixture models (GMM) [2,16], hidden Markov models (HMM) [7,14], support vector machines (SVM) [17], random forests [3] or different deep neural network (NN) architectures [4,9,10,12]) have been applied to this field but they have been mostly evaluated on publicly available benchmark datasets, which are composed of audio scenes from everyday life (e.g. TUT sound event databases [15] or DCASE databases [13]). Up to now, there are no datasets which specialise on sound events from manufacturing environments. Especially, the task of connector locking detection has been hitherto comparatively little explored [1].

For this reason, this paper investigates the performance of different neural network architectures in order to distinguish different connector locking events from each other respectively from other "fake" events with similar acoustic properties in a manufacturing environment. Due to the short sound duration of connector plugging events, for the first time, an optical membrane-free sensor with high sampling rates is used for the acoustic data acquisition. In order to increase the robustness of the connector locking assessment, multimodal sensor data are extracted in order to improve the classification accuracy. In particular, in addition to sound data, video data obtained by capturing the workflow during the execution of manufacturing working tasks is used. In this paper, a neural network based framework for assessment of connector plugging is presented. The processing of sound and video data is discussed in detail and the performance of all different network architectures is analyzed.

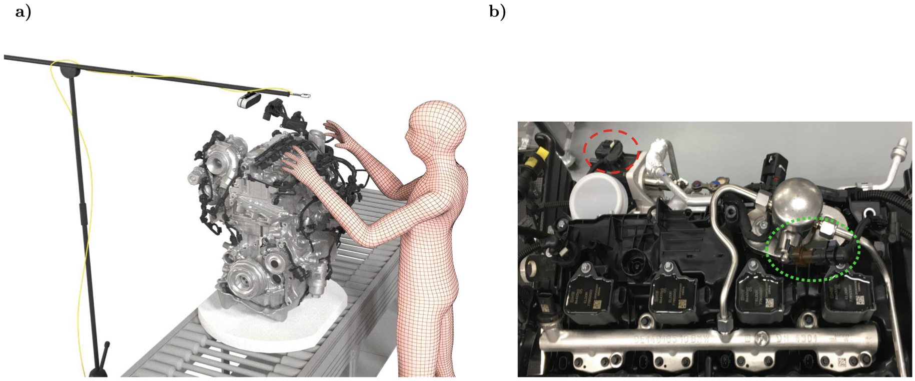

Fig. 1. (a) Experimental setup for multimodal data acquisition. (b) The analyzed connector types 1 and 2 are highlighted with a green and a red circle. (Color figure online)

2 Optical Microphones

The working principle of the used membrane-free optical microphone exploits the properties of a rigid Fabry-Pérot interferometer [8]. Laser light is transmitted through an optical fibre to a semi-reflective two-mirror system within the sensor head. Sound events from the external environment can penetrate the etalon through a small aperture window. Subsequently, the changing pressure of the sound wave causes fluctuations of the refractive index within the laser propagating medium, i.e. the change of the refractive index is leading to slight wavelength shifts of the transmitted and reflected laser light in the Fabry-Pérot interferometer. In order to draw conclusion on the occurring sound event, the outgoing laser light from the etalon interferes with the incoming laser beam which is leading to detectable laser intensity fluctuations. These intensity fluctuations are transformed to analogue voltage signals which can be used for the examination of the acoustic signal.

In contrast to state-of-the-art membrane microphones, optical microphones offer linear frequency responses from 10 Hz up to 1 MHz and allow sampling rates of up to 4 MHz. Thus, it is possible to resolve sound events with time durations below 1 ms (e.g. connector locking events) very accurately.

3 Sound Data Acquisition and Representation

3.1 Data Generation

Since there are no publicly available datasets for the task of connector lock detection, it is mandatory to acquire a sufficiently large set of locking sound data in order to train a supervised machine learning model. To this end, the optical microphone has been installed at the final assembly line of an engine manufacturing plant. At the considered workstation, electric connectors must be plugged into engine blocks. In order to avoid collisions of engine parts with the sensor head, the optical microphone had to be placed at a distance of approximately 50 cm away from the connector locking position.

At the analyzed workstation, two different connector types are plugged which both comprise a primary and a secondary lock. A connector is correctly plugged when the primary and secondary lock are pushed into the corresponding socket. Consequently, a correct plugging is characterized by two click events with a well-defined temporal separation. The working contents of the analyzed workplace do not only include plugging but also other working processes (e.g. screwing). Thus, it is a main aim of this investigation to not only be able to distinguish between two different connector types but also to separate primary lock and secondary lock events from background events and sound events with similar acoustic properties (i.e. fake events). The experimental setup as well as the analyzed connector types are depicted in Fig. 1.

The start and stop of the sound measurements have been triggered with digital outputs whenever a new engine is conveyed to the workstation. In order to assign the acoustic signal to the corresponding sound source, video files of the working tasks have been captured simultaneously for each acoustic sound sample taken.

3.2 Data Annotation

In order to train a supervised machine learning model, the generated sound data have to be annotated. Thus, it is mandatory to determine at which point in time the locking events occur. For this reason, those local maxima from the analogue signal need to be determined, whose signal-to-noise ratio exceeds a predefined threshold level A_{thres} and whose minimal temporal separation lies above a threshold t_{sep}.

Within the proposed analysis the datasets have been labeled by hand, i.e. the determined maxima at the given time instances are classified according to the executed work step recorded on the corresponding video files. Within the proposed analysis the following classes have been considered: background (BG), primary lock connector 1 (CP1), secondary lock connector 1 (CS1), primary lock connector 2 (CP2), secondary lock connector 2 (CS2), fake event (FE). The background events are generated from recorded data before the locking process of connector 1 has been initialized. Those maxima which could not be assigned to a connector locking event are classified as fake events. Subsequently, in sum 1,223 data samples have been generated, of which 988 samples have been used for training, while the remaining 235 data samples are used for testing. The distribution of data classes is given in Table 1.

Table 1. Distribution of data classes

Classes	Distribution [%]
Background (BG)	26
Primary lock connector 1 (CP1)	15
Secondary lock connector 1 (CS1)	16
Primary lock connector 2 (CP2)	15
Secondary lock connector 2 (CS2)	10
Fake event (FE)	18

3.3 Feature Generation

Before training the model, the feature representation that is best suited for the classification task should be generated from the input data. The information about a sound event class is typically not stored in a single time frame but over a consecutive temporal context, i.e. the feature representation comprises a temporal sequence of feature vectors. The analyzed connector locking sound events typically last for approximately 1 ms. In order to avoid a cropping of the locking sound signal, a maximal feature time duration of 3 ms is chosen (1 ms before and 2 ms after the sound peak maximum). By applying a sample rate of 4 MHz a time window of 3 ms corresponds to a feature length n_F of 12,000 for the extracted analogue signal. In order to determine the optimal feature length, n_F is treated as a hyperparameter and optimized by means of a grid search. Thus, n_F highly depends on the feature space chosen, i.e. in frequency space the feature length is set together by the number of frequency contributions times the analyzed time steps.

Since the use of optical microphone data for audio event detection is so far an unexplored discipline, three different types of input features from time and frequency domain are investigated in terms of their locking event detection performance:

a) *Analogue time signal*: Due to the high sampling rate of the optical microphone, the time signal can be resolved very accurately. The feature representation of the connector locking event is described by 12,000 input features.
b) *Log-STFT signal*: From the logarithmic frequency spectrum of the Short-time Fourier Transform (STFT) follows, that primary and secondary locking events show frequency contributions up to 100 kHz, which correspond to the first 40 amplitude contributions from the log-STFT spectrum. In total, this gives a log-STFT feature length n_F = 6,000.
c) *MFCC features*: In the field of sound event detection the frequency information from the STFT is further processed in order to find an optimal set of sound event features. In many cases the spectogram is transformed to the mel scale, which shows higher frequency resolutions in lower frequency domains. By applying Discrete Cosine Transforms over the mel spectogram, Mel Frequency Cepstral Coefficients (MFCC) can be generated which are often used

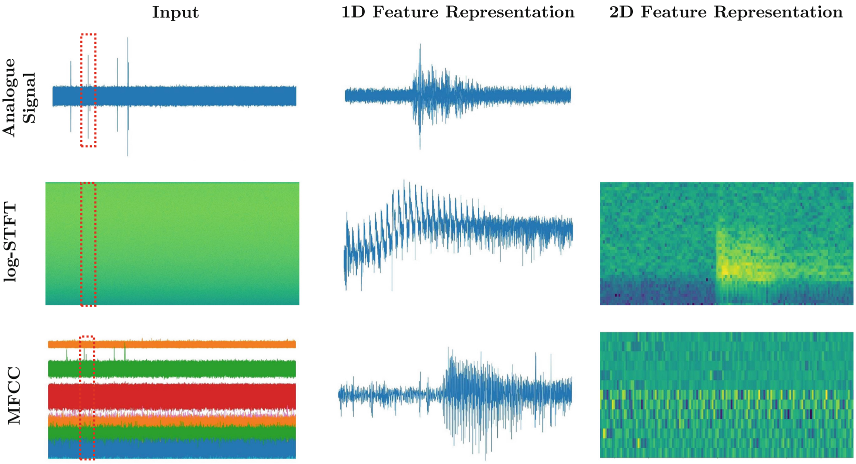

Fig. 2. Comparison of input data and feature representations of a locking event generated from analogue signal, log-STFT signal and MFCC features. The locking events are highlighted in the input data with a red dashed box. (Color figure online)

for the spectral representation of acoustic signals [6]. Within this paper the first 13 MFCCs are considered which correspond to an overall MFCC feature length of 1,950.

All of the investigated features are normalized to a range [−1,1]. In contrast to the other feature representations analyzed, the MFCC features are standardized with mean zero and a standard deviation of one before getting normalized. The feature representations generated on the basis of their input signals are exemplified in Fig. 2.

4 Data Processing

Since manufacturing environments are well-known for machinery noise, it is questionable whether the correct locking of connector cables can be determined solely from the acoustic properties of clicking events. Thus, in the proposed work the performance of unimodal as well as multimodal neural network architectures are both investigated for the specific task of detecting connector locking events. The different types of neural network architectures analyzed are introduced in the following.

4.1 Unimodal Neural Network Architectures

Neural network architectures tend to outperform other approaches for sound event detection, and the following four different architecture types are considered in the proposed work.

a) *Neural Network*: The chosen feed-forward neural network architecture (NN) is composed of multiple hidden layers followed by the output layer used for classification.
b) *1D-Convolutional Neural Network*: One-dimensional convolutional neural networks (CNN) are investigated in order to capture temporal correlation effects of the input features. The CNN architecture is set up by multiple convolutional hidden layers followed by pooling layers.
c) *2D-Convolutional Neural Network*: Two-dimensional CNN architectures are the state-of-the-art approach for image data classification, and the log-STFT spectrum as well as the MFCC features are not only analyzed with a one-dimensional but also with a two-dimensional feature representation. Consecutive 2D convolutional layers are applied in connection with max pooling layers, which are fed into fully connected layers, before being transferred to the final classification layer.
d) *Combined Model (NN+CNN)*: In order to make use of the acoustic signal information in time and frequency domain, a combined model is introduced, which makes use of a NN for analogue signal features, while the frequency features are fed into a one-dimensional CNN branch. Both networks are merged into two joint fully connected layers, which are followed by the output layer for classification.

Table 2. Hyperparameter choices for neural network architecture optimization

Hyperparameter	Range
Feature length n_F fraction	1/10, 1/6, 1/2, 1
Number of layers	1, 2, 4, 6
Number of neurons per layer and input feature length	1, 2, 4, 6
Learning rate	0.01, 0.001, 0.0001

All of the proposed architectures are analyzed for different sets of hyperparameters. A grid search is carried out for all network types in order to find the optimal choice of hyperparameters. The range of used hyperparameters for the grid search are given in Table 2.

4.2 Multimodal Neural Network Architectures

In order to further improve the validation performance of the described neural network architectures for connector locking event classification, it is beneficial to extend the amount of gathered information from the plugging processes. Further, using multiple input sources is beneficial in order to better distinguish and characterize similar manufacturing working steps. These different feature representations can all be fed individually into separate branches of a multimodal

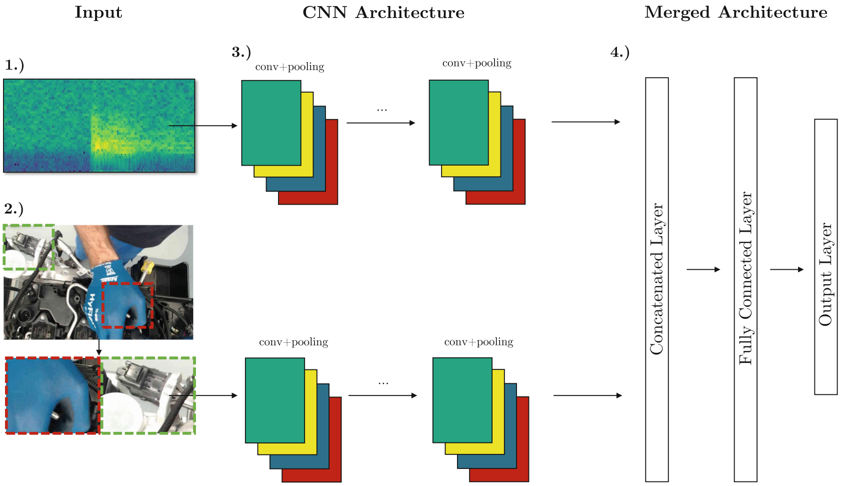

Fig. 3. Multimodal neural network architecture using log-STFT features as input.

neural network architecture. As an example the processing steps of a multimodal architecture are shown for log-STFT features in Fig. 3 and explained in the following.

1.) For the proposed task the multimodal neural network processes the sound features in the first branch of the network - the audio branch.
2.) The second branch - the visual branch - processes the recorded video files during process execution. As the plugging processes are carried out at the same location for all engine types, small image patches ($100 \times 100 \times 3$) centered at the two connector locking positions are cropped from the video files and merged into one image patch ($100 \times 200 \times 3$). These images are chosen from the video file in accordance with the instance in time when an acoustic peak is determined.
3.) The visual branch consists of a 2D-CNN architecture which is composed of several blocks of convolutional and pooling layers that are followed by fully connected layers.
4.) The last fully connected layer of the visual branch is merged with the last fully connected layer of the audio branch. The concatenated layer is again fed into fully connector layers followed by the output layer which is then used for classification.

5 Experimental Results

The trained models for connector lock event detection are validated on the basis of the following two evaluation metrics: *Accuracy* and the F_1*-score*. The accuracy is calculated as

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

with TP, TN, FP, FN corresponding to the number of true positive, true negative, false positive and false negative predictions. The F_1-score is determined as

$$F_1 = 2 \cdot \frac{PRE \cdot REC}{PRE + REC} \tag{2}$$

with PRE being the classification precision and REC the classification recall.

The overall results of the unimodal (sound only) and multimodal (sound and visual) neural network architectures are given in Table 3. With only reaching maximal accuracy values close to 75% (for the log-STFT with 2D-CNN architecture), one can deduce, that the sole use of acoustic information processed by the analyzed neural network architectures is not robust enough to classify connector locking events in a manufacturing environment. From the results follow that log-STFT and MFCC feature representations are clearly preferable over the sampled analogue signal for the task of connector locking classification. With regard to the investigated neural network architectures, 2D-CNNs exceed the performance of all other architectures.

Table 3. Accuracy and F_1 results obtained for connector locking detection using the proposed unimodal and multimodal neural network architectures.

Method	Unimodal architectures						Multimodal architectures			
	Analogue Signal		log-STFT		MFCC		log-STFT		MFCC	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
NN	0.54	0.48	0.64	0.55	0.66	0.62	0.63	0.61	0.72	0.68
CNN (1D)	0.62	0.57	0.66	0.61	0.72	0.69	0.74	0.71	0.84	0.82
CNN (2D)	-	-	0.75	0.71	0.69	0.62	0.84	0.81	0.84	0.82
NN + CNN	-	-	0.66	0.61	0.70	0.64	-	-	-	-

Compared to the unimodal results, the investigated multimodal approaches outperform all investigated unimodal approaches with maximal accuracy levels close to 85% (for the log-STFT/MFCC and 2D-CNN architecture). Thus, exploiting multiple process information can help to describe complex tasks like the correct locking of connector cables. A more detailed evaluation of the results (illustrated by the confusion matrix of Fig. 4) shows, that the occurring error can be mainly attributed to mispredictions of primary and secondary locking events for both connector types. Apart from these false predictions, there are still a few cases where fake events get predicted as locking events. This scenario is definitively more problematic than connector lock mispredictions, because in the worst case the plugging process would be classified as correctly executed,

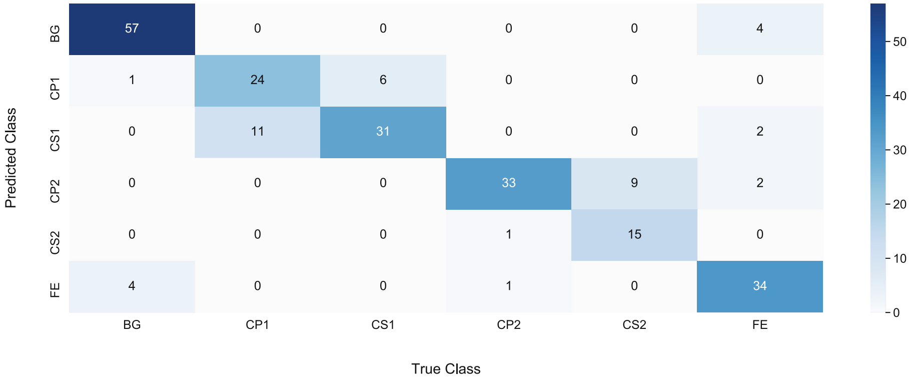

Fig. 4. Confusion matrix for the evaluation of the 235 test data samples using the best performing log-STFT multimodal network architecture.

although the connector might not even be plugged at all. In practice this case would only occur, when two consecutive fake events are wrongly classified as primary and secondary lock events of a specific connector type. Nevertheless, since most connector plugging processes are not validated at all at the assembly lines, the proposed connector lock detection framework can lead to a significant improvement of the quality validation in manufacturing. By integrating the framework in series production, additional data can be collected and the model can be further optimized.

6 Conclusion

In this paper, the task of connector locking detection has been investigated by making use of optical microphones. Different unimodal and multimodal neural network architectures have been trained in order to distinguish primary and secondary locking events of two different connector types from events with similar sound characteristics respectively from each other and hence to assess whether connectors were correctly plugged. The obtained results indicate that multimodal neural network architectures making use of acoustic and visual process information clearly outperform unimodal approaches that only take into account sound features and achieve connector locking classification accuracy scores close to 85%. Since currently there is no check for correct connector plugging, the proposed framework can help directly to increase the quality of process execution in manufacturing.

It would be of great interest to analyze, if additional contextual process information from sensor data (e.g. the force or pressure measured at the thumb during the plugging process) could help to eliminate the occurring mispredictions and would thereby allow a more robust use in manufacturing. Furthermore, the used laser microphone had to be positioned at a comparatively high separation distance which is accompanied by an attenuation of high-frequency components and thus partly annihilate the advantage of the investigated optical sensor. Instead, one could install small membrane microphones in the glove of an employee and thereby extract sound events better which only occur in the close vicinity of the microphone. Thus, the acquired sound data could potentially lead to a more robust detection of connector lock events in combination with the proposed multimodal framework.

References

1. Aoyagi, M., Ueno, T., Okuda, M.: Automatic detection system for complete connection of a waterproof soft-shell electronic connector with a sliding locking device. IEEE Sens. J. **9**(3), 285–292 (2009). https://doi.org/10.1109/JSEN.2008.2012225
2. Atrey, P.K., Maddage, N.C., Kankanhalli, M.S.: Audio based event detection for multimedia surveillance. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 5, pp. 813–816, May 2006. https://doi.org/10.1109/ICASSP.2006.1661400
3. Barchiesi, D., Giannoulis, D., Stowell, D., Plumbley, M.D.: Acoustic scene classification: classifying environments from the sounds they produce. IEEE Signal Process. Mag. **32**(3), 16–34 (2015). https://doi.org/10.1109/MSP.2014.2326181
4. Cakir, E., Heittola, T., Huttunen, H., Virtanen, T.: Polyphonic sound event detection using multi label deep neural networks. In: 2015 International Joint Conference on Neural Networks, IJCNN, pp. 1–7, July 2015. https://doi.org/10.1109/IJCNN.2015.7280624
5. Cho, H., Kim, Y., Kim, B., Song, J.: A strategy for connector assembly using impedance control for industrial robots. In: 2012 12th International Conference on Control, Automation and Systems, pp. 1433–1435, October 2012
6. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980). https://doi.org/10.1109/TASSP.1980.1163420
7. Eronen, A.J., et al.: Audio-based context recognition. IEEE Trans. Audio Speech Lang. Process. **14**(1), 321–329 (2006). https://doi.org/10.1109/TSA.2005.854103
8. Fischer, B.: Optical microphone hears ultrasound. Nat. Photonics **10**, 356–358 (2016). https://doi.org/10.1038/nphoton.2016.95
9. Hayashi, T., Watanabe, S., Toda, T., Hori, T., Le Roux, J., Takeda, K.: Duration-controlled LSTM for polyphonic sound event detection. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(11), 2059–2070 (2017). https://doi.org/10.1109/TASLP.2017.2740002
10. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 131–135, March 2017. https://doi.org/10.1109/ICASSP.2017.7952132
11. Jorg, S., Langwald, J., Stelter, J., Hirzinger, G., Natale, C.: Flexible robot-assembly using a multi-sensory approach. In: Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation Symposia Proceedings, vol. 4, pp. 3687–3694, April 2000. https://doi.org/10.1109/ROBOT.2000.845306. (Cat. No.00CH37065)

12. Marchi, E., Vesperini, F., Eyben, F., Squartini, S., Schuller, B.: A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 1996–2000, April 2015. https://doi.org/10.1109/ICASSP.2015.7178320
13. Mesaros, A., et al.: DCASE 2017 challenge setup: tasks, datasets and baseline system. In: DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events. Munich, Germany, November 2017. https://hal.inria.fr/hal-01627981
14. Mesaros, A., Heittola, T., Eronen, A., Virtanen, T.: Acoustic event detection in real life recordings. In: 2010 18th European Signal Processing Conference, pp. 1267–1271, August 2010
15. Mesaros, A., Heittola, T., Virtanen, T.: TUT database for acoustic scene classification and sound event detection. In: 2016 24th European Signal Processing Conference, EUSIPCO, pp. 1128–1132, August 2016. https://doi.org/10.1109/EUSIPCO.2016.7760424
16. Oldoni, D., De Coensel, B., Rademaker, M., De Baets, B., Botteldooren, D.: Context-dependent environmental sound monitoring using SOM coupled with LEGION. In: The 2010 International Joint Conference on Neural Networks, IJCNN, pp. 1–8, July 2010. https://doi.org/10.1109/IJCNN.2010.5596977
17. Rakotomamonjy, A., Gasso, G.: Histogram of gradients of time-frequency representations for audio scene classification. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 142–153 (2015). https://doi.org/10.1109/TASLP.2014.2375575