Acoustic validation of snap fit assembly

Egge Rouwhorst
S377196

July 5, 2023

Supervisors:	*Affiliation:*
Dr. A.O. Krushynska	RUG
Dr. A.G.P. Kottapalli	RUG
Ing. R. Kristiaan	FMI ImProvia

Contents

1 Introduction

In industry some assembling steps consist of attaching one part to another via a snap fit. When done by hand, validation of the connection is inherent in the process due auditory, visual or tactile feedback. If the assembly is robotized validation becomes necessary to ensure correct manufacturing of the product. This project is focused on snap fits in shaver components from Philips. Previous work on this topic was done by Kasper Hendriks at FMI ImProvia using tactile feedback, which showed promising results. However, the results were not satisfactory and an acoustic approach was desired instead because of the no-contact measurement possibilities. This report explores the feasibility of validating a snap fit using acoustics. In the next section the design goals are stated together with the system boundaries.

1.1 Design boundary and goals

The system contents are given in table 1. Endogenous variables describe the system, where the system design and the variable influence each other. Exogenous variables are ones that affect the design of the system but cannot be changed to fit the system. The last column features excluded variables which were not considered in the system.

Endogenous	**Exogenous**	**Excluded**
Number of microphones for the sensor	Noise levels in the factory	The effect of temperature or humidity on operation.
Number of snap fits that can be validated simultaneously	Number of snap fits that are connected simultaneously	Assembly of snap fits of another process
The source of power for the system	The type of flaw that occurs	
The method to detect flaws		
The method to indicate flaws		
Distance between the microphone and snap fit		

Table 1: System boundaries.

The acoustic sensor should have the following criteria:

1. Differentiate between “good” and “bad” snap fits.
   The sensor should have a precision, recall and true negative rate of 99.5% for the good snap fit.

2. Relatively cheap initial and operational costs.
   The sensor should not cost more than €1000 in initial costs and no more than €400 annually.

3. Robustness in operation:

    (a) Perform in a noisy, industrial environmental.
    The sensor will operate in an assembly line at Philips during typical factory operation within specified criteria.

    (b) Run 24/7.
    The sensor should be able to run continuously without intervention, given the sensor is well maintained.

    (c) Handle small product variations.
    The sensor should be able to discern flaws in a snap fit, even with small variation in the components.

    (d) Handle multiple types of products.
    The sensor should be able to discern flaws in a snap fit for different parts without a hardware-based change.

And lastly the stakeholders for this project are FMI ImProvia, Philips and the customer of Philips or FMI ImProvia. For FMI ImProvia and Philips the interest is high, and only FMI ImProvia also has high influence.

1.2 Literature

No articles specifically on the detection of flaws in a snap fit with sound have been found. So a more broad search has been applied. There do exist studies on snap fit validation using force. One such study uses machine learning and a mixed human and robot assembly [1]. Others looks at features in the force profile [2] or forces in the robotic arm itself, supplemented with machine learning [3]. Fault detection is an emergent field and an ever growing number of papers and reviews exist on the topic [4, 5, 6]. Acoustic based monitoring is also increasing in interest with are multiple names for more or less the same subject, namely acoustic anomaly detection (AAD), anomaly detection in sound (ADS) and anomalous sound event detection (anomalous SED). More broad is the term subsequence outliers in time series [7]. Various papers exist outlining the challenges and possible solutions from detection and equipment setup to various methods of signal analysis, anomaly detection and data classification [8, 9]. Relevant papers overview classification methods and machine learning models and are often linked with databases to test the anomaly detection, e.g. the ones available from the DCASE 2020 event [10, 11, 12]. All of the mentioned articles use some form of machine learning to detect anomalies, however, the sound data usually spans a longer time period of seconds to minutes, instead of milliseconds such as from a snap-fit. An overview of machine learning approaches often used in literature can be seen in Figure 1.1. Feature extraction and selection are the front-end processing, which determine which data is put into the machine learning model. Multiple types of feature extraction can be used simultaneously.

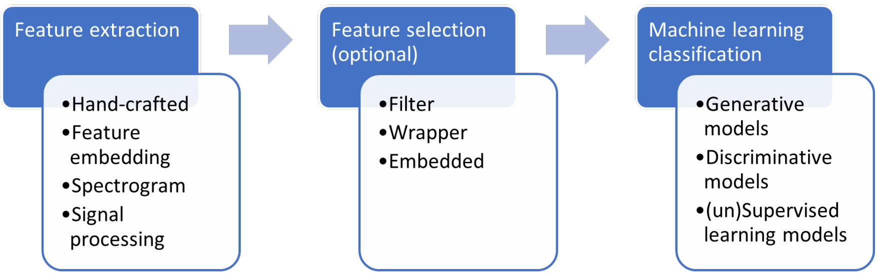


Figure 1.1: Overview of common machine learning approaches for anomaly detection.

2 Design choices

2.1 Snap fit

Since this is a proof-of-concept design, the type of snap fit could be chosen. At FMI ImProvia multiple different snap fits related to a shaver were available. From these only one was supposed to be reversible, which is a snap fit used in the attachment of a shaver front plate to a rack used in a spray paint line. Although the snapping sound was weaker for this attachment than compared with some others, the reversible design was preferred as many tests are desired. In the other cases the quality of the joints would degrade quickly or many shaver parts would be required. In figure 2.1 the snap fit in question can be seen. To keep the hook part in a steady horizontal position a holder was created, which can also be seen. Wrong snaps were introduced as a missing hook (snap 3) and a deformed hook (snap 6).

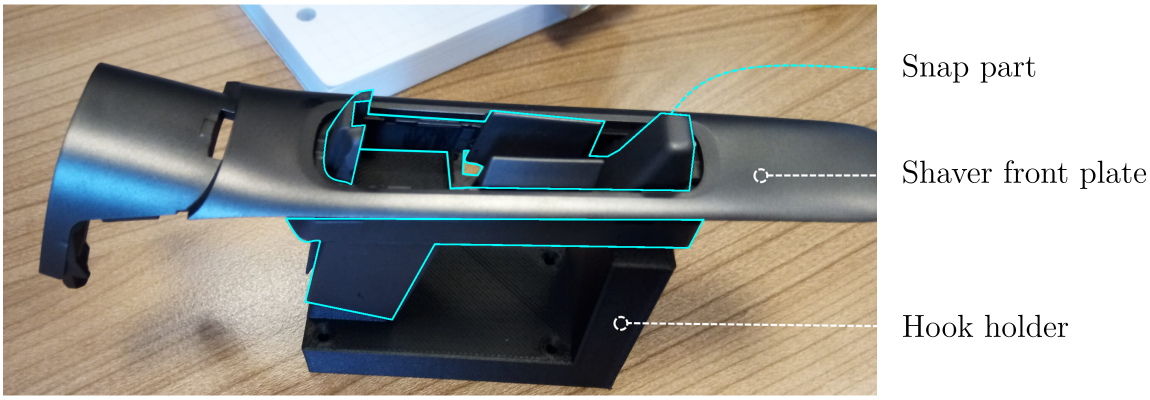


Figure 2.1: Shaver front plate and the two-sided hook in the middle on top of the hook holder.

2.2 Microphone

The microphone ideally has a good signal to noise ratio for both internal noise and external noise. The external noise rejection methods considered were: using a directional microphone, using multiple microphones and minimizing the distance between microphone and snap. Directional microphones can quickly become expensive and do not filter out the noise coming from the direction they are pointed in. This means they have to be used in conjunction with a sound-absorbing backdrop and/or multiple microphones. Sound absorbing materials generally only work in some range and adding more microphones drives up the cost even more. Measurements of the noise background at Philips showed levels reasonably below the sound of a snap at 10 cm, as can be seen in figure 2.2. This means that omnidirectional microphones would also work when positioned close enough to the source. Since a snap sound has a very short duration – the main envelope has a duration of up to 10 ms – high temporal resolution is especially beneficial. For this a microphone with a high frequency range, coupled together with a fast analogue to digital converter (ADC) would be required. A micro-electromechanical systems (MEMS) microphone is suited for this. These microphones are very cheap, with a general price of around €2 per piece. The microphone chosen was the Infineon IM73A135V01, as seen in figure 2.3. The ADC used was the Motu M2, which measured with a sample rate of 192 kHz and is purchasable for around €235 [13], although a sufficient ADC can be bought for less. The microphones required a typical power of 2.75 V which was supplied using a lab power supply. The measurements were performed at very close range and in the case of two microphones, these were employed on either side of the snap fit.

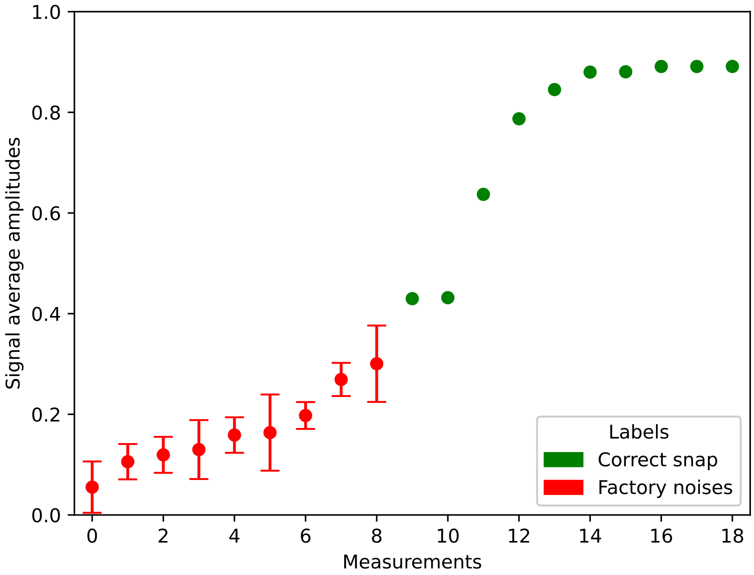


Figure 2.2: Average signal amplitude for both noise and snaps. Noise average amplitude is taken as the average of the maxima of 1000-sample frames.

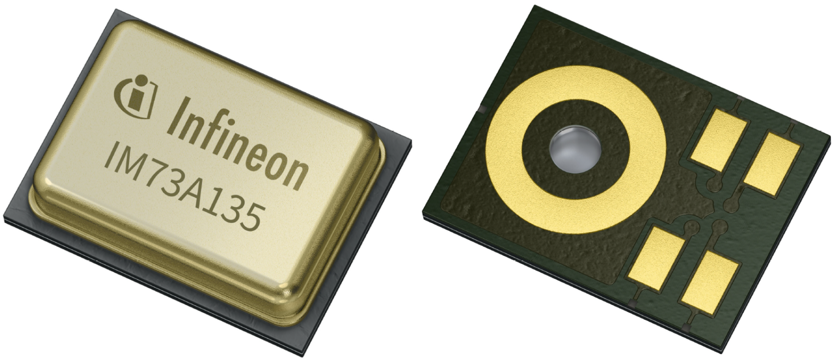


Figure 2.3: MEMS microphone top and bottom [14].

2.3 Actuator

For consistency of the snap fit mating an actuator from Festo Automation was used. Since design of a specific gripper to undo the snap fit was out of the scope of this project, a simple attachment was applied to a linear motor. The attachment consists of four legs, pressing on the corners around the snap fit to ensure reasonably uniform pressure, as can be seen in figure 2.4.

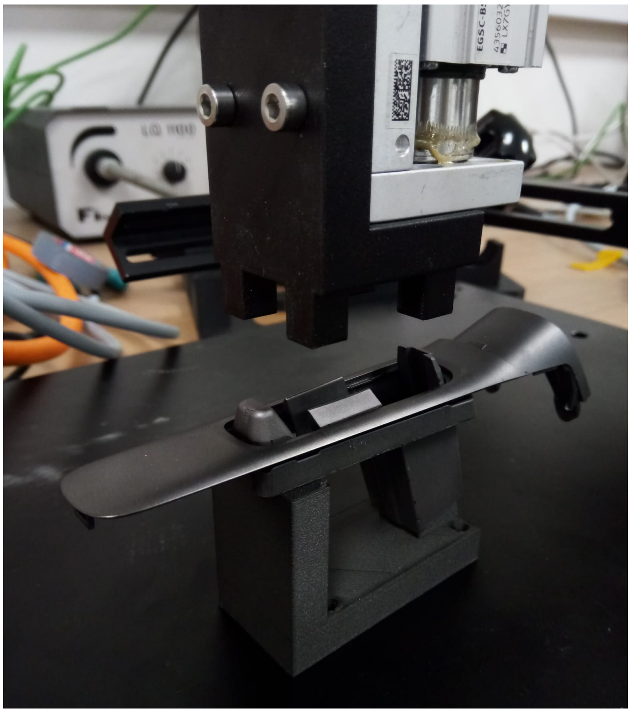


Figure 2.4: Linear actuator with attachment for snapping the front plate into place.

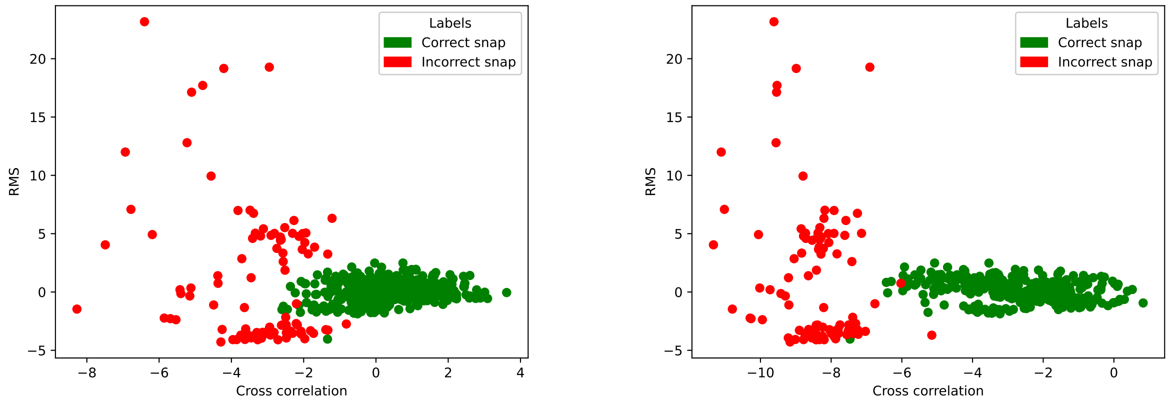
2.4 Data acquisition & storage

The data was gathered in the following way: first the setup was configured for the *session*, such as the position of the microphone and the hook holder. Then multiple measurement *sets* were performed in which 25 to 35 times the actuator pushed down a chosen front plate, which was manually undone. During each whole set the computer collected data from microphone via Audacity®. Afterwards the snaps were extracted from the big audio file to small ones by selecting a up to 8000 samples around a peak to form a so-called frame. This last step is done to make each snap easily accessible for labelling and training. For convenience and speed these small audio files are also used for testing, although in principle also the full session audio files could be used with a sliding window method.

2.5 Feature extraction

The features can be classified into temporal and spectral features. Many features have been looked at for classification but for brevity only a couple will be explained here. Two important temporal features are the cross-correlation of the audio sample (frame) with a training set and the root mean square (RMS) of the sample. Cross-correlation is used as it is a very powerful tool in comparing sequential data. It can be described as a normalized sliding dot product, from this sliding dot product the time step with the highest value is selected and is taken to be the correlation between the two audio vectors (or matrices). This sliding dot product is necessary as the signals are not aligned by the pre-processing step. To compare two collections of vectors (train and test sets), a matrix of these cross correlations needs to be constructed. To get a single feature per audio vector this matrix needs to be collapsed down into a vector. This can be done by using e.g. the 0.85'th percentile or the mean of the 5 top correlating entries for each test sample. The advantage of using the percentile seems to be that it works better with the local outlier factor (LOF) classifier due to a higher spread in the training set, but is not as robust as using the mean of the top 5 as the latter is not sensitive to adding different types of snaps, see figure 2.5, which means the classifier is less prone to false positives.

A spectral feature implemented is a short-time Fourier transform on a sliding window of 512 samples, a step size of 64 samples and unless specified otherwise the window function used was the Hann window. The windows are combined together into a spectrogram for the whole frame. The cross-correlation of this spectrogram is the final feature. Spectral (or rather cepstral as it's called for Mel) features such as Mel filter banks and their energies and derivatives, as suggested by literature [11, 15], were found to be lacking in performance. In general, the spectral features do not seem to contain large contrast between good and bad snaps, which might be due to the very short and impulse-like sound, which do inherently have a low frequency resolution.



(a) Cross correlation with 0.85 percentile setting and RMS.

(b) Cross correlation with highest 5 setting and RMS.

Figure 2.5: The features from the single MEMS microphone session with different choices for the cross-correlation. In both cases good and bad are reasonably disjoint but the mean of the top 5 correlations creates a larger separation. The average of the training set however ($x = 0$) is much higher, causing the LOF to classify many good snaps as wrong.

In table 2 an overview of features investigated is gives, where they have been ranked by sequential feature elimination to maximize the $F_{0.2}$ score when training on the dataset from figure 3.5.

2.6 Classification

The OCSVM classifier attempted first did not show a great fit, partially as it assumed the data contained some incorrect snap samples. Hence a different classifier was employed called ‘local outlier factor’ (LOF), with ‘novelty detection’ enabled which means the model can be trained on good snaps only. The reason to train the classifier on good snaps only is because the actual failures that can occur can be very diverse and rare. A representative failure database to train on is therefore not feasible to obtain. This density-based classifier showed very good results. One downside with the local outlier factor was that when the training set increased, the decision boundary seemed to become stricter, as the model became even more sensitive to variation. To alleviate this, a custom classifier could be created which sets the decision boundary for the temporal cross correlation at e.g. either 30% or at 4 sigma from the training mean, depending on which is the highest. This classifier was found to have an equal to slightly better performance in one situation, as can be seen in 3.1. The downside of such a custom classifier is of course that it does not generalize well to more features or different situations and hence for ease of use the LOF was chosen as classifier and care was taken to make the training set not too large.

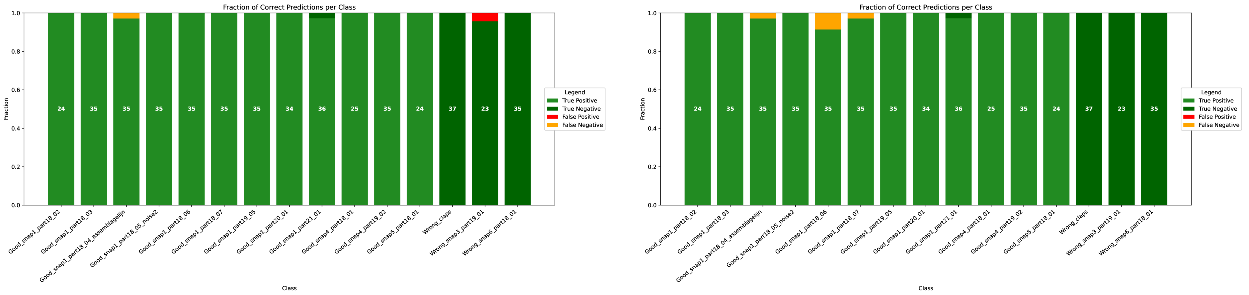
3 Results & Discussion

3.1 Preliminaries

Preliminary measurements were done with a in the factory to gather various samples from the environment. With the smartphone as microphone also 45 good snaps were performed where the actuator was set to 0.4 cm/s and a bit of tape was applied to the interface, both to minimize the sound from the attachment striking the shaver front plate. With a 30/70 train/test split, the model with only the raw signal cross correlation performed flawlessly on the test set. The test set was duplicated five times and each one was distorted by adding noise from a different sample from the factory. From these 5 times 32 snaps sounds, only one was flagged as wrong, given the signal was fed through a high pass filter first, indicating a strong capability for noise rejection. A minor dataset of five situations, of which two good and three wrong, containing three snaps each was also measured with the phone. This minor dataset was classified as all wrong by the former model. This is most likely due to the small changes in the test setup and indicates the model is actually quite specific.

3.2 Follow up

Measurements with the MEMS microphone have been performed and comparable results as the preliminary measurements were found, see figure 3.1. Classification in a single measurement run using cross-correlation on the raw data shows incredibly high accuracy. However, when e.g. the hook component or the front plate was swapped, the cross correlation between the two measurements runs dropped significantly, from 69% to 59% on average, indicating differentiation capabilities. Because of this, if the decision boundary is set relatively high the classification will largely exclude these different, yet correct, snaps. Setting the boundary too low, e.g. lower than 30%, causes inclusion of incorrect snaps. The accuracy of the classification in this case depends on the classifier and the variety of training data. Because the model already showed precision and recall of over 99.5% in some situations, the main focus moving forward was on investigating the robustness to product variations.



(a) Custom classifier.

(b) LOF classifier.

Figure 3.1: Model performance using raw signal correlation and RMS features with two different classifiers. The training data consists of 30% of snap 1, snap 4 and snap 5 with part 18 measurement runs.

3.3 Binaural recording

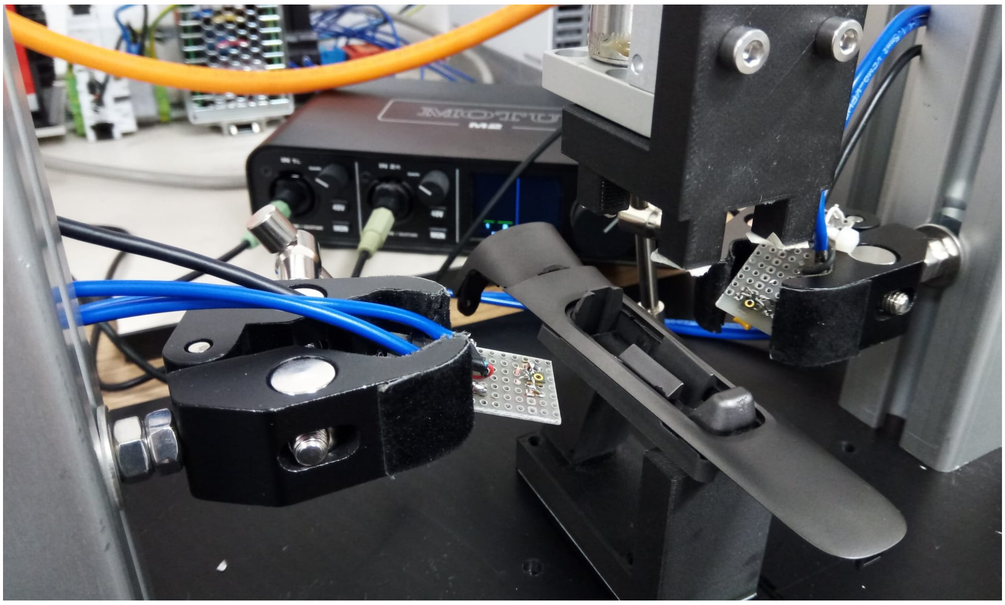
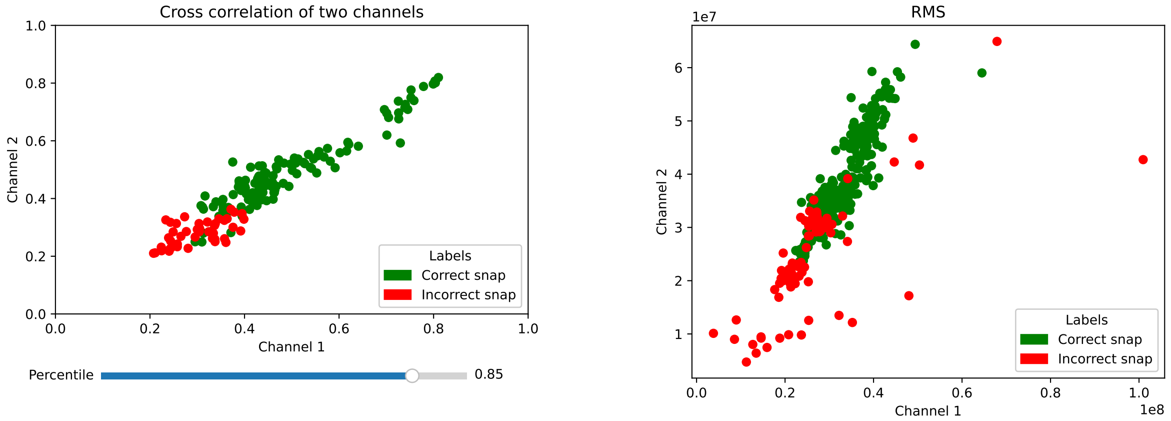


Figure 3.2: Binaural measuring with the robuster setup.

A more robust version of the setup was created and can be seen in figure 3.2. Measuring with two microphones showed both microphones pick up roughly the same signal envelope, however, the correlation between the two raw signals is quite low at around 30%. This is most likely due to the different location. Since they pick up roughly the same signal, we expect the features to be linearly related. Indeed, in figure 3.3 can be seen that this is the case. The incorrect snap outliers are due to non-snap sounds, such as a clap or undoing a snap fit. The linear relation implies that the symmetric placement of the microphones does not drastically improve distinction between good and bad clicks. Investigation of this by comparing the classification of either channel alone versus using both channels indeed shows the classification of dual channel to be somewhat better: the $F_{0.2}$ score improved from 0.996 to 0.997 and for the later test from 0.983 or 0.956 to 0.987. The improvement being more significant for one channel indicates that the position of the sensor most likely also plays some role, which can be reduced by using more microphones.



(a) Correlation features.

(b) RMS features.

Figure 3.3: Binaural feature plots, showing a linear relation between the two recording channels. The off-diagonal incorrect outliers in the RMS features are unrelated sounds such as undoing the snap fit.

3.3.1 Product variation

To investigate the effect of the product variation, six different snap components have been tested, all with the same front plate. It was found that the signals differed significantly between the measurement runs. In figure 3.4 the smoothed absolute signals of the different measurement sets can be seen. As can be seen snap 9 and snap 10 (red and blue lines) look similar but differ from the rest. The same is the case for the light green and orange lines. Lastly note that the signal from snap 7, which was measured twice, differed significantly. It appears the process of placing the hook on the holder as well as the hook part itself cause major variation in the signal. Another experiment with a different hook holder, as seen in figure A.1, shows the same effects in addition to variation within some measurement, see figure 3.5. This high variability between measurement sets means that training on only a couple of measurement sets causes false negatives for sets not included. The other measurement runs can be added to the training data to account for the misclassification of these runs as was done before in figure 3.1. However, there are a couple downsides with that approach: firstly, the amount of processing required for prediction increases as more cross-correlations will need to be computed and secondly, there is no guarantee that this will eventually be robust to future variations in the products or setup. Moreover, if the runs that are added contain too much variation or are similar to incorrect snaps, the classifier can become too broad and cause false positives.

From figure 3.5 it can also be seen that a bad snap (snap 3, for which one hook was removed) can have a very similar envelope to a good snap, e.g. snap 7 part 18 run 4. This similarity means that if only snap 7 is used for training it will more easily classify snap 3 as good, while snap 9 for example will be classified as bad. And indeed, only training on snap 7 has as result that almost primarily only snap 7 is recognized even when many features are included. Of course, it might be the case that there exists some feature that belongs exclusively and necessarily to a good snap in general but if it exists it is very hard to find. Secondly, there is no salient difference in features between the two channels for snap 3, indicating that this double microphone setup does not clearly differentiate between the location of the snap for this type of snap fit. This latter observation is most likely due to the small distance between the hooks, and multi-channel recording might still be useful for other snap fits where the hooks have a larger spacing.

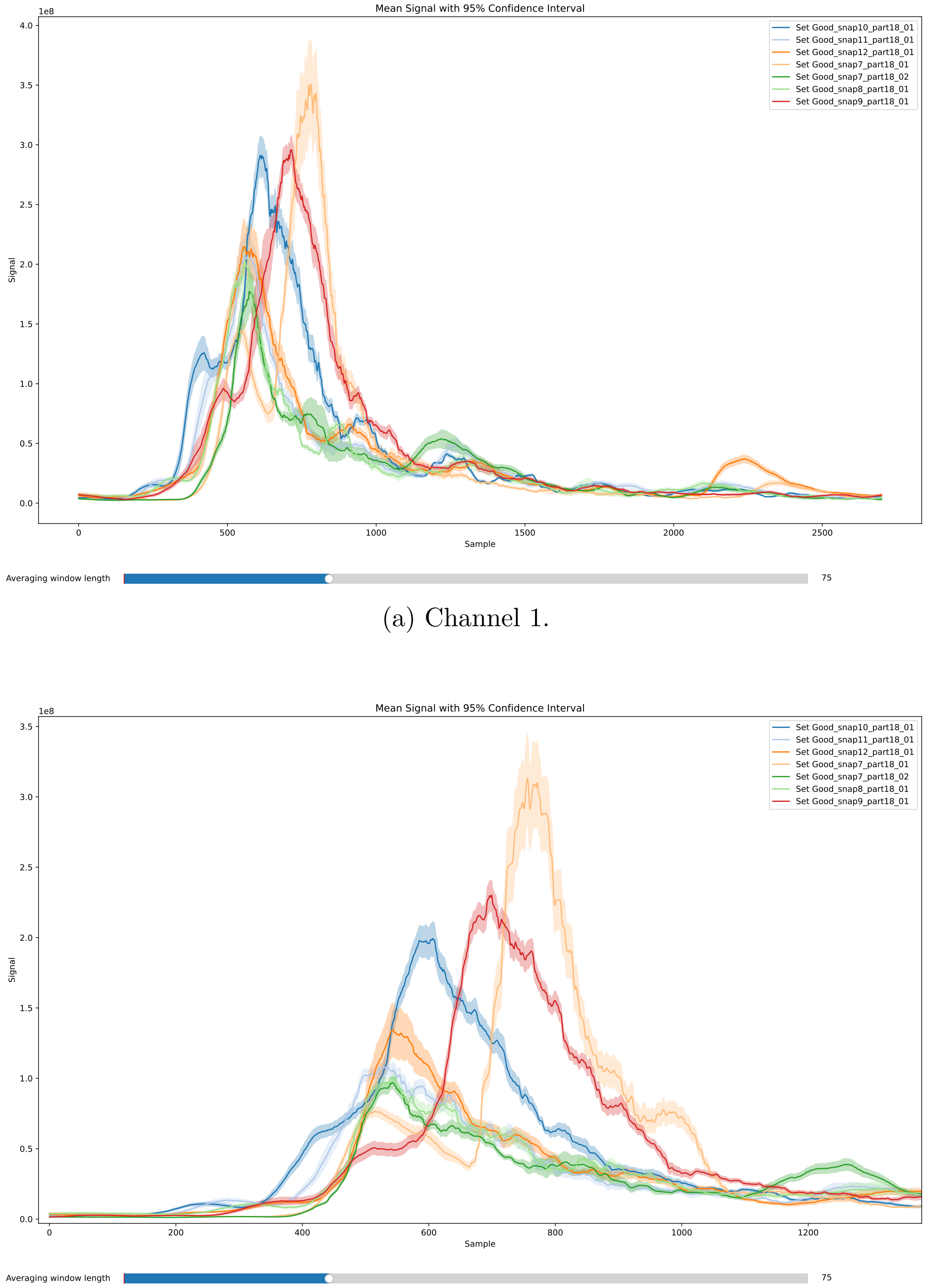
4 Conclusion

Whereas initially the classification seemed highly successful, robust snap-fit sound classification was frustrated by large variation of the signal due to (re)placement of the hook components. High accuracy is mainly only obtained when the model is trained and tested on measurement sets where the hook component has not been replaced. Since components are expected to be replaced often, or even only are supposed to snap once, training on exactly the same components as the test set is not possible. If acoustic snap-fit validation is to be used in industry the variation in the snap sounds should be brought to a minimum to successfully find outliers. The number of microphones and the amount of snap fits that can be validated simultaneously were both determined to be one, in the specific snap fit of this study.

Regarding the design criteria, the sensor does in come conditions have a remarkably high precision and recall of over 99.5%. The true negative rate reached 98.9% due to a lack of negatives. If however only a small set of hook components can be used for the test set, some these requirements are not met. The setup was very cheap and hence under budget due to the microphones being very cheap. From the few tests the noise rejection of both the microphone and the model was found to be very good. The robustness to small product variations is where the sensor fails. The sensor can in principle also be used for different products, given the location of the microphones is fixed or refixable.

5 Acknowledgements

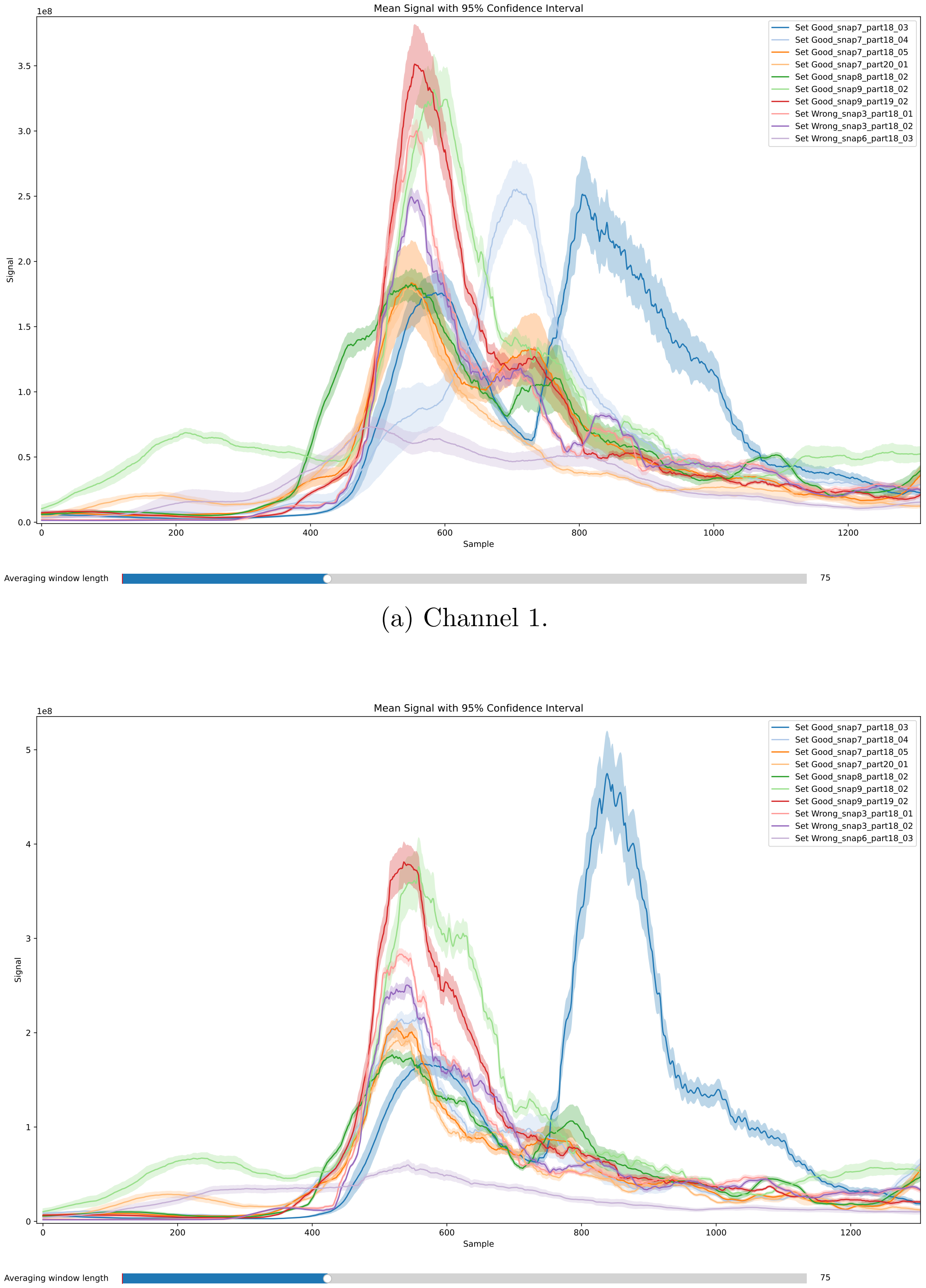
I would like to thank my academic supervisor A. Krushynska for the input and meetings. Secondly, I am grateful for my colleagues at FMI ImProvia for assisting me continuously throughout the project.



(a) Channel 1.

(b) Channel 2.

Figure 3.4: The audio envelopes of multiple measurement sets. The averaging window length is set at 75 points.



(a) Channel 1.

(b) Channel 2.

Figure 3.5: The audio envelopes of multiple measurement sets with a different hook part holder. The averaging window length was set at 75 points.

References

[1] S. Doltsinis, M. Krestenitis, and Z. Doulgeri, “A machine learning framework for real-time identification of successful snap-fit assemblies,” *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 1, pp. 513–523, Jan. 2020. [Online]. Available: https://doi.org/10.1109/tase.2019.2932834

[2] F. Radil, R. Adámek, B. Dobossy, and P. Krejčí, “Robotic snap-fit assembly with success identification based on force feedback,” in *Modelling and Simulation for Autonomous Systems*. Springer International Publishing, 2022, pp. 145–157. [Online]. Available: https://doi.org/10.1007/978-3-030-98260-7_9

[3] J. Rojas, K. Harada, H. Onda, N. Yamanobe, E. Yoshida, K. Nagata, and Y. Kawai, “Towards snap sensing,” *International Journal of Mechatronics and Automation*, vol. 3, no. 2, p. 69, 2013. [Online]. Available: https://doi.org/10.1504/ijma.2013.053409

[4] B. Peng, Y. Bi, B. Xue, M. Zhang, and S. Wan, “A survey on fault diagnosis of rolling bearings,” *Algorithms*, vol. 15, no. 10, p. 347, Sep. 2022. [Online]. Available: https://doi.org/10.3390/a15100347

[5] J. Chen, C. Lin, D. Peng, and H. Ge, “Fault diagnosis of rotating machinery: A review and bibliometric analysis,” *IEEE Access*, vol. 8, pp. 224 985–225 003, 2020. [Online]. Available: https://doi.org/10.1109/access.2020.3043743

[6] A. Abid, M. T. Khan, and J. Iqbal, “A review on fault detection and diagnosis techniques: basics and beyond,” *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3639–3664, Nov. 2020. [Online]. Available: https://doi.org/10.1007/s10462-020-09934-2

[7] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, “A review on outlier/anomaly detection in time series data,” *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–33, Apr. 2021. [Online]. Available: https://doi.org/10.1145/3444690

[8] G. Jombo and Y. Zhang, “Acoustic-based machine condition monitoring—methods and challenges,” *Eng*, vol. 4, no. 1, pp. 47–79, Jan. 2023. [Online]. Available: https://doi.org/10.3390/eng4010004

[9] M. Jones, D. Nikovski, M. Imamura, and T. Hirata, “Anomaly detection in real-valued multidimensional time series,” in *2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY conference, Stanford university, may 27-31, 2014*, Jun. 2014. [Online]. Available: https://www.merl.com/publications/docs/TR2014-042.pdf

[10] E. C. Nunes, “Anomalous sound detection with machine learning: A systematic review,” 2021. [Online]. Available: https://arxiv.org/abs/2102.07820

[11] Z. Mnasri, S. Rovetta, and F. Masulli, “Anomalous sound event detection: A survey of machine learning based methods and applications,” *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5537–5586, Dec. 2021. [Online]. Available: https://doi.org/10.1007/s11042-021-11817-9

[12] Y. Wang, Y. Zheng, Y. Zhang, Y. Xie, S. Xu, Y. Hu, and L. He, “Unsupervised anomalous sound detection for machine condition monitoring using classification-based methods,” *Applied Sciences*, vol. 11, no. 23, p. 11128, Nov. 2021. [Online]. Available: https://doi.org/10.3390/app112311128

[13] Bax-shop, motu m2. [Online]. Available: https://www.bax-shop.nl/externe-audio-interfaces/motu-m2-audio-interface

[14] *IM73A135V01*, 2021st ed. [Online]. Available: https://eu.mouser.com/datasheet/2/196/Infineon_IM73A135_DataSheet_v01_00_EN-3163938.pdf

[15] E. Kiktova, M. Lojka, J. Juhar, and A. Cizmar, “Comparison of feature selection algorithms for acoustic event detection system,” in *Proceedings ELMAR-2014*. IEEE, Sep. 2014. [Online]. Available: https://doi.org/10.1109/elmar.2014.6923312

Appendices

A Pictures

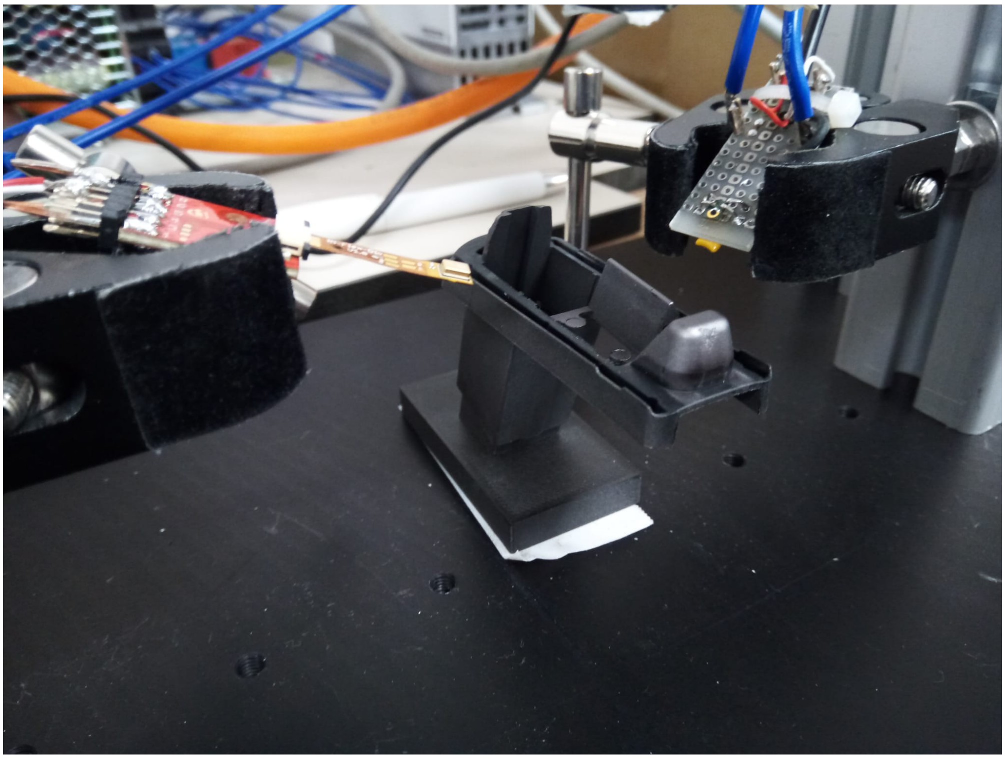


Figure A.1: Replacement of the hook part holder.

B Feature list

Feature name	Ranking set 1	Ranking set 2
Cross correlation	6	17
Spectral coherence	12	14
Spectral correlation	9	
Spectral correlation, Blackman-Harris	7	
Spectral correlation, Nutall	<6	
Spectral correlation, Taylor	8	
Spectral centroid	10	<6
Spectral flatness	11	16
Spectral polyfit		15
Spectral brightness	<6	6
Spectral peak frequency	<6	<6
Spectral centroid maximum	<6	
RMS	<6	11
Flux		<6
Loudness		13
Spectral peaks		8
Distribution		7
Entropy		<6
Crest		12
Strong decay		10
Attack time		9
Attack steepness		<6

Table 2: Feature ranking list. The notation <6 implies the feature is in the top 5.