

# Multimodal Object Analysis with Auditory and Tactile Sensing using Recurrent Neural Networks

Yannick Jonetzko<sup>1</sup>, Niklas Fiedler<sup>1</sup>, Manfred Eppe<sup>2</sup>, and Jianwei Zhang<sup>1</sup>

<sup>1</sup> Technical Aspects of Multimodal Systems

<sup>2</sup> Knowledge Technology

Universität Hamburg, Vogt-Kölln-Straße 30, 22527 Hamburg

{jonetzko, 5fiedler, eppe, zhang}@informatik.uni-hamburg.de \*

**Abstract.** Robots are usually equipped with many different sensors that need to be integrated. While most research is focused on the integration of vision with other senses, we successfully integrate tactile and auditory sensor data from a complex robotic system. Herein, we train and evaluate a neural network for the classification of the content of eight optically identical medicine containers. To investigate the relevance of the tactile modality in classification under realistic conditions, we apply different noise levels to the audio data. Our results show significantly higher robustness to acoustic noise with the combined multimodal network than with the unimodal audio based counterpart.

**Keywords:** Multimodal · Neural Network · Tactile · audio · Object Analysis

## 1 Introduction

In a domestic environment designed by and for humans, robots have to overcome similar challenges and conditions. Therefore, the employed sensors are often inspired by humanoid senses, mimicking their form and functionality. For successful interaction and exploration, it is essential for the robot to rely on the quality and accuracy of its perception. Nevertheless, in diverse natural scenarios, not all senses are reliable or can be used at all, due to environmental conditions, such as ambient light or acoustic noise. To overcome these handicaps, humans use several senses combined. In the field of robotics, the advantages of using multimodal input are extensively explored. However, the majority of the work focuses on combining visual information with other modalities (e. g., [1,2,3,4,5]). In contrast, our work is focused on the integration of acoustic and tactile data. In Sec.

---

\* This research was funded by the German Research Foundation (DFG) and the National Science Foundation of China in project Crossmodal Learning, TRR-169. It is also partially financed by the H2020-MSCA-RISE Project ULTRACEPT. Manfred Eppe acknowledges support via the DFG-funded IDEAS (EP 143/2-1) and LeCAREbot (EP 143/4-1) projects.

2 of this article, we describe existing unimodal auditory and tactile approaches for object analysis and classification. We also describe multimodal approaches that use vision in combination with additional sensors. However, to the best of our knowledge, there exists no approach that systematically investigates in how far acoustic noise can be addressed by integrating auditory and tactile information. To address this gap, we present a novel approach that performs well in a scenario where i) visual data is not useful, ii) the audio signal is noisy, and iii) tactile data is available but not as accurate as the other senses. This motivates our following research question:

*To what extent can the multimodal analysis of noisy audio data and tactile information lead to a significant improvement of the classification accuracy of bulk materials in visually indistinguishable objects?*

To address this question, we perform experiments with an anthropomorphic robotic hand. The hand is equipped with bio-inspired tactile sensors and shakes multiple visually indistinguishable objects in front of a microphone.<sup>3</sup> Each object contains different pills which the robot classifies using audio and tactile information (see Sec. 3). We preprocess the signals using Mel coefficients (see Sec. 4) and perform classification with recurrent neural networks (see Sec. 5). Our evaluative comparison of the multimodal network with the individual counterparts (see Sec. 6) provides details on the performance gain when combining auditory and tactile data. Sec. 7 concludes the results.

## 2 Related Work

Our multimodal approach builds on audio and tactile measurements to classify pills in identical containers. The following approaches explore related problems.

### 2.1 Tactile object analysis and classification

Tactile measurements are applied in various applications, such as braille reading [6] or fall detection of elderly people based on vibration measurements on the ground [7]. We use the BioTac sensor for tactile measurements. This sensor was initially presented by Wettels et al. [8] and has since been used for various tasks including tumor localization [9], force estimation, and slip detection [10].

We use the sensor to measure vibrations that result from exploratory movements. Fishel and Loeb [11] demonstrate that the device “exceeds human performance in detecting sustained vibrations”. Xu et al. [12] elaborate on active tactile perception by proposing six exploratory movements that cause tactile sensations: pressure, lateral sliding, static contact, enclosure, hefting, and contour following. The first three are supposed to generate sensory information with a BioTac sensor, while the others rely on joint positions and forces of the robot.

<sup>3</sup> To enable control over the acoustic noise, we used a high-quality external microphone and added separately recorded noise of the robot to the signal during the evaluation.

Specifically, lateral sliding was used to generate vibration data as one modality for object classification. Another approach that focuses explicitly on texture classification by lateral sliding is the work by Kerzel et al. [13], who achieve more than 99% classification accuracy for 32 different materials.

In the field of classifying container contents based on tactile data with machine learning, the work of Chen et al.[14] is probably most related to our approach. The authors perform multiple experiments based on the vibration signatures, which they acquire while shaking containers filled with various objects. In their object classification experiments, the authors were able to classify shaken objects into one of 12 classes with an accuracy of  $93.8 \pm 4\%$  using a Support Vector Machine. In contrast to our approach, they used a custom shaking mechanism that involves a contact microphone mounted at the container. Therefore, the resolution of the measurements is higher, and the signal is most likely less noisy than in our work. Additionally, the object types are more diverse than the ones classified in this work (e. g., ball bearing, acrylic piece, rubber ball).

## 2.2 Acoustic object analysis and classification

Durst and Krotkov [15] classify objects based on the sound resulting from an impact between an aluminum cane and the object itself. A Fourier transform was applied for preprocessing. The performance of a minimum-distance classifier was compared to a decision-map classifier, which combines a minimum-distance with a decision-tree classifier. They reached comparable results for classifying objects into five different classes with an accuracy of 94.2% and 93.8%, respectively. Luo et al. [16] also use the sound of an impact between the object and a marker pen to classify the object. They create a dataset by knocking 120 times with the pen on 30 different objects. Similar to our work, they feed the resulting signal into a Deep Neural Network. Therefore, the authors use stacked denoising autoencoders to train the network and reach a classification accuracy of 91.50%.

Our presented audio processing is inspired by the work of Eppe et al.[17]. The authors present an approach where the content of visually indistinguishable plastic capsules is determined by a robot that analyzes an audio signal resulting from shaking the capsules. The material type is classified and the weight estimated using the resulting audio samples. More diverse object classes are used in their work, while our approach distinguishes a range of pill types. In their case, the shaking movement is faster, so that the produced sound is generated from the material hitting the capsule walls. In contrast, because of hardware constraints, we slowly rotate the container, which produces sliding and rattling sounds when the pills slowly slide down the container wall.

Liang et al. [18] use audio signals to analyze robotic pouring and estimate the liquid height in several containers. In this scenario, it is hard to estimate the height with visual sensors because of the transparency of liquids.

### 2.3 Multimodal object analysis and classification

The multimodal object categorization approach proposed by Nakamura et al. [1] investigates unsupervised learning of object categories using visual, audio, and tactile measurements fed into an adaption of probabilistic Latent Semantic Analysis (pLSA). Though the authors consider unsupervised categorization, whereas our work concerns supervised object classification, their findings are very relevant to ours, as they demonstrate that combining audio and tactile data leads to improved categorization results.

Sinapov et al. [19] present another example of object categorization based on multimodal perception. The authors categorize 36 unique containers (three container colors, four content types, and three content weights) with a robot by learning their features while performing ten exploratory movements. In addition to the categorization of objects into groups, their approach is able to learn relations of object pairs and groups. Parts of their results show that shake and rattle movements are well suited for content and weight detection via audio and proprioception measurements.

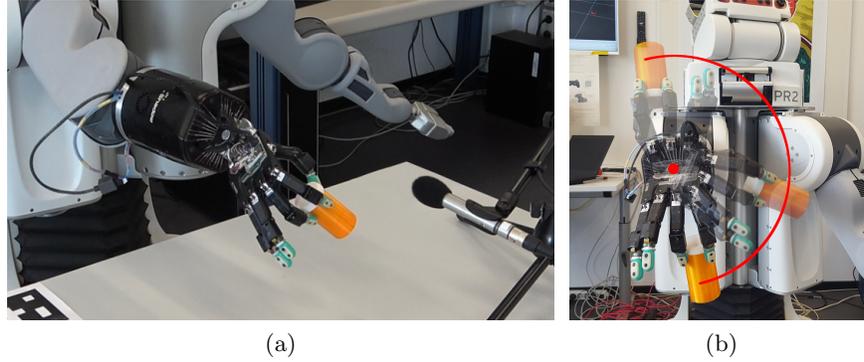
Pieropan et al. [2] propose a “method for audio-visual recognition of human manipulation actions”. In their work, the authors classify human manipulation actions (e. g., opening a milk bottle or pouring cereal) based on visual and audio data. For the classification, audio features, orientations, and positions of manipulated objects are fed into a Hidden Markov Model (HMM). Similar to our work, the audio features are extracted by a Mel Frequency Cepstral Coefficients (MFCC) analysis of the raw signal. The authors are able to classify the actions with an average accuracy of 73% over seven classes (including the class *no action*).

## 3 Experiment Setup

In the experimental setup, a PR2 robot [20] shakes eight 3D printed medicine containers that are visually indistinguishable. Container contents that the robot is supposed to classify are listed in Tab. 1. The robot holds a container with

**Table 1.** Pill classes included in the data set and used in the experiment

									
		Magnesium	Calcium	B-Complex	Big Mints	Chew	Small Mints	Vitamin B	Candy
	Weight per pill	1.27g	2.2g	0.31g	1.2g	1.13g	0.5g	0.55g	0.6g
Sample count	One pill	166	99	104	110	118	137	126	137
	Small amount	228	405	222	212	299	218	260	391
	Half full	239	407	232	413	251	293	290	336
	Full	237	184	242	215	243	296	304	174
	Overall	820	1095	800	950	911	944	980	1008



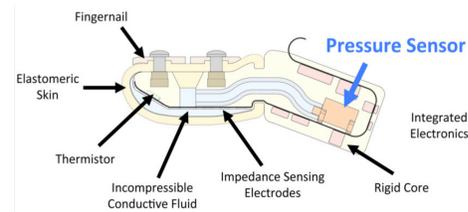
**Fig. 1.** Overview of the experiment setup. **(a)** The Shadow Dexterous Hand holds a medicine container filled with pills to classify with three fingers. BioTac sensors are mounted at each fingertip of the hand. An external microphone is placed in front of the hand to record the resulting noise. **(b)** Depiction of the applied shake motion. Each container was rotated up- and downwards again along the red line with a rotation velocity of 0.8 and 1 radian per second.

a tripod grasp, using three tactile-sensitive fingertips. The content is classified based on auditory and tactile measurements acquired while shaking the container in front of a microphone (see Fig. 1a).

### 3.1 Sensors

For the experiment, the PR2 robot is equipped with a five-finger 19 degrees of freedom Shadow Dexterous Hand [21] (see Fig. 1a). Each of the five fingers is fitted with a tactile BioTac sensor [22] at the last phalanx. The sensor mimics the human fingertip and its sensory modalities, as depicted in the cross-section of the sensor in Fig. 2.

The bone-like rigid core is surrounded by a flexible silicone hull filled with a conductive liquid. With its various embedded sensors, it is capable of measuring multiple modalities, like pressure, temperature, and the deformation of the silicon hull. A hydroacoustic pressure sensor at the end of a small tube inside the rigid core of the BioTac measures the pressure of the liquid, which changes during contacts. To gather the absolute fluid pressure (DC), the transducer output is amplified with a gain of 10 and low-pass



**Fig. 2.** A cross-section of the BioTac sensor used for the acquisition of vibration measurements. The pressure sensor used in this work is marked in **blue**. [11]

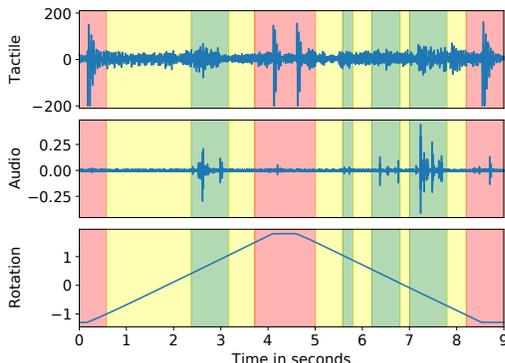
filtered with 1040 Hz. Afterward, the signal is band-pass filtered between 10 and 1040 Hz with a gain of 99.1 and produces the dynamic fluid pressure (AC) value, also referred to as vibrations. 19 electrodes, distributed over the rigid core, measure changes in impedance due to the dispersion of the conductive liquid. Based on this information, the deformation of the hull can be inferred. The resulting spacial tactile information is not useful for the classification of the bottle’s ingredients as the containers are physically similar. The liquid inside the BioTac is indistinguishable from water regarding the vibration properties [11]. For the classification evaluated in this work, we consider only the AC value because neither the temperature nor the hull deformation provides any meaningful information to classify the pills. Since no direct contact with the pills is made, we also do not consider thermal conductivity, as proposed by Xu et al. [12]. In addition to the tactile information, an audio signal is used for the classification. As shown in Fig. 1a, a microphone is pointing to the sound source of the rattling pills. We record a single 44.1 kHz acoustic signal, which is later downsampled, as described in Sec. 4.

### 3.2 Pill Container

In our considered scenario, the robot can only rely on its tactile and audio perception because the containers are visually equal. To ensure the same preconditions for each type, the containers were 3D printed in a typical medicine container optic, with an FDM printer and PLA plastic. We chose the size of the can in a way that it is easy to grasp for the robot, with an inner height of 80 mm, a radius of 20 mm, and a wall thickness of 2.5 mm.

### 3.3 Dataset Recording

We recorded the dataset in a robot laboratory with ordinary environmental noise (e. g., computers running, nearby airport). The robot’s fans, motors, and gears produce most of the noise. To achieve comparable auditory and tactile data, the robot arm is straightened out, pointing forward to the microphone and rotates around the rotatory joint between elbow and forearm. Before the shaking movement, an experimenter hands over the medicine container and firmly places it inside a tripod grasp with the thumb, the fore- and the middle finger. Only the tactile fingertips touch the object to reduce vibration noise produced by the robot in the tactile readings. In each iteration, the force applied to the object, as well as the finger poses, remain constant to produce homogeneous data. The actual shaking motion is shown in Fig. 1b. One shake is either a lifting or lowering 180 degree rotation around the middle axis of the forearm. Each of the 8 pill classes is recorded with 2 different velocities (0.8 and 1.0 radian per second) and 4 different amounts of pills, with 12 shaking motions a time. This results in an overall number of 768 shaking movements in the dataset. Each sample includes information about the pill class, the number of pills, the number of shaking movements, and the angular velocity. Besides the audio and tactile data, the joint position of the forearm is recorded.



**Fig. 3.** Exemplary data stream from 10 Big Mint pills shaken at an angular velocity of 0.8 radian per second. The tactile and audio signals and the applied rotation of two shaking processes are shown. Parts of the stream which are used for training and testing are marked in **green**. Areas marked in **red** are excluded from the data set based on the rotation of the hand, while those marked in **yellow** are removed because the audio signal did not surpass a threshold defining a minimum signal strength.

## 4 Data Preprocessing

To maximize the amount of relevant information in the data, we extract the samples with features of particular interest. That is, we cut samples from the auditory and tactile data stream based on the rotation angle and the maximal audio signal amplitude, as described in Sec. 4.1 and Fig. 3.

The preprocessing of the selected samples involves *Mel Frequency Cepstral Coefficients* (MFCC) [23], a common method to preprocess human speech and other auditory data. MFCC have been successfully applied among others in the field of music synthesis [24] and acoustic classification [17,25]. Since tactile data can also be decomposed into different frequencies, we apply MFCC to both tactile and audio data (see Sec. 4.2 and Sec. 4.3). Preliminary experiments have demonstrated that applying MFCC to tactile information provide a significant performance gain.

### 4.1 Sample Selection

The realistic experiment setup leads to much noise and unusable samples in the dataset, due to ego noises of the robot, in particular, the fans and the motors. Fig. 3 depicts the raw data stream of two shaking motions, starting with a lifting movement that ends after 4.5 seconds, followed by a returning motion (see Fig. 1b). To select those samples of the signal that contain useful information about the container contents, we apply two filter stages consecutively. The first filter stage (marked red in Fig. 3) removes the samples close to the direction change outside the joint position range of  $-1$  to  $1.5$  radians. Due to the jerk (see Rotation in Fig. 3), especially the tactile signal is very noisy in that region and

consequently useless for classification. The second filter stage (marked yellow in the diagram) considers that the pills inside the container do not produce sounds and vibrations throughout the whole shaking process, which results in many samples without significant features. These silent samples are removed from the dataset by considering the maximal amplitude of sound signal frames with a size of 0.2 seconds. Samples below a threshold are filtered out as they merely contain noise. We assume that all other samples contain useful information about sound and vibration, the pills inside the containers produce (highlighted with green).

## 4.2 Auditory Data

The MFCC feature extraction process can be split into multiple stages. After a Fourier transformation is applied to sequences of the audio signal, the resulting power spectrum is mapped onto the Mel scale. First, logarithm activation is applied to the frequency energies, and second, a discrete cosine transformation is used, which results in a spectrum of the amplitudes, the MFCC.

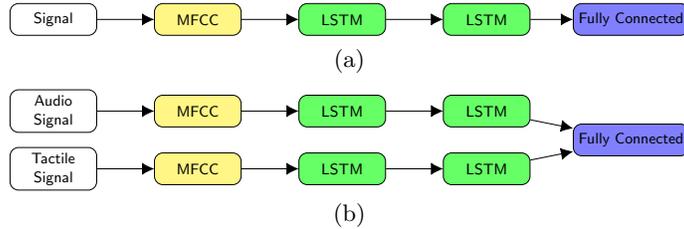
Multiple parameters are important for the calculation of MFCC. In our experiments, we considered the following: window size, step size, and amount of Mel coefficients. All of these parameters were determined with Tree-Parzen-based hyperparameter optimization [26], using a hyperparameter space similar to that of Eppe et al. [17]. We achieved the best results with a window size of 0.03, a step size of 0.02, and 21 Mel coefficients.

## 4.3 Tactile Data

The tactile data for each finger was recorded at a rate of 1000 Hz. Due to the high frequency, it is possible to use methods from audio processing for the tactile signal. Similar to the auditory data, we applied MFCC to the normalized tactile signal, further explained in the previous paragraph. Experiments showed improvements using the MFCC over feeding the raw tactile signals directly into the neural network. Since the sample rate of the tactile sensor is significantly lower compared to a microphone, the range of the analyzed frequency spectrum has to be adapted accordingly. Again, the same parameters as for the audio data, as well as the frequency range, were tested with hyperparameter optimization. We achieved the best results with a window size of 0.04, a step size of 0.04, 9 Mel coefficients, and a frequency range from 4 Hz to 440 Hz.

## 5 Neural Network Architecture

The architecture of the neural networks is based on the system proposed by Eppe et al. [17], but extended with the tactile modality. To this end, we define two separate neural networks, one for the classification of tactile measurements and another one for auditory data. While the general architecture of both networks is similar (see Fig. 4 a), they differ mainly in the MFCC parameters, as mentioned before, and parameters of the specific layers.



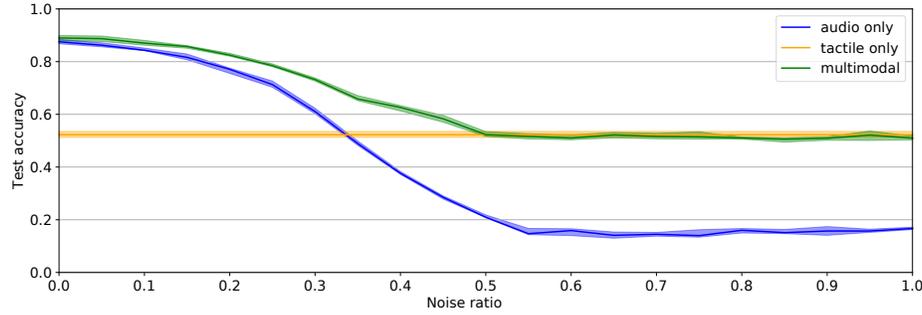
**Fig. 4.** Schematic view of the (a) unimodal and multimodal (b) neural network architectures used in this work.

We feed features extracted by the MFCC into two consecutive *Long Short-Term Memory* (LSTM) layers [27] which use a RELU activation function. The last hidden LSTM state of each data sequence is piped into a fully connected layer with dropout. Then, to perform the classification, we apply a softmax activation. Through hyperparameter optimization [26], the parameters of the layers were optimized for each modality separately. For the audio-based classifier, we achieved the best results with 400 units in the first LSTM, 90 units in the second one, and a dropout rate of 0.34. In the tactile version, 180 and 90 units worked best in the first and second LSTM, respectively. A dropout of 0.7 leads to the highest classification accuracy. The usage of *Gated Recurrent Units* (GRU) [28] instead of LSTMs was evaluated but did not impact the classification accuracy of the architectures.

To use multimodal input in the classifier, the developed architecture had to be adapted for two input sequences with differing sample rates and signal frequency range. Therefore, the multimodal version of the architecture concatenates the unimodal networks by feeding the results of the recurrent layers into one fully connected layer with eight neurons (see Fig. 4 b). Due to sampling rates and MFCC parameters, the sequence sizes differ. The fully connected layer for the multimodal case is configured similarly to the one in the unimodal architecture. We trained the resulting architecture with an Adam optimizer [29] at a learning rate of  $lr = 0.001$  and the default momentum parameters provided by the Keras deep learning framework, i. e.,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ . In preliminary tests, we also considered variations of our architecture, in particular, the application of adding a second dense layer for both modalities combined as well as one additional dense layer for each. However, we have not observed any significant effect on the classification results.

## 6 Results

For the training and evaluation of our system, we split the dataset into 80% training and 20% testing samples. An overview of the classification results is provided in Tab. 2. The best classification accuracy on the testing split of the dataset was 56.06% for tactile only data, 89.1% on audio only data, and 91.23% with multimodal input. To have optimal control over the amount of noise, the



**Fig. 5.** Comparison of the classification accuracy of both unimodal and the multimodal approaches with increasing noise ratio in the audio signal. trained and tested ten times. The mean of the results of each step consisting of ten train and test iterations is drawn as a solid line as well as the 25- and 75-percentiles are indicated translucently.

microphone was placed at a position which allowed to record the audio signal as clearly as possible (see Fig. 1a). In practical applications, where the microphone is mounted directly to the robot, more noise is to be expected. To explore the effect of ego-noise under controllable conditions, multiple training and test runs were conducted in which the prerecorded noise of the robot was added to the original audio signal. The classification accuracy of the developed networks was evaluated with regard to the noise ratio by training and testing individual audio and multimodal networks in 21 different noise conditions with a noise ratio between 0 and 1. Noise in the audio signal does not affect the results of the network designed for the tactile input signal. Therefore, only one evaluation step was carried out with the tactile architecture. In each evaluation step, ten iterations of training and testing were conducted. Fig. 5 shows the results of the study as a plot of the mean accuracy of each network, as well as a depiction of the 25- and 75-percentiles.

For a noise ratio of 0.3, the confusion matrices for the three models are given in Tab. 2. The matrices link the actual pill type (rows) and the classified ones (columns), with the corresponding success rate. In Tab. 2 a the performance of the unimodal audio network is visualized, the accuracy of this sample is 58.75%. The overall classification accuracy of the given median tactile model is 51.25% (see Tab. 2 b). While segment (a) shows distributed weaknesses in the detection independently of the pill type, matrix (b) indicates a clear accumulation of detection insecurities. In contrast, the multimodal-based network shows a significantly better accuracy of 71.63% compared to the unimodal counterparts (see Tab. 2 c). None of the confusion matrices indicate a strong within-pair confusion, meaning that none of the models performed particularly badly in distinguishing one class from another.

**Table 2.** Confusion matrices for all three models at a noise ratio of 0.3

	Magnesium	Calcium	B-Complex	Big Mints	Chew	Small Mints	Vitamin B	Candy		Magnesium	Calcium	B-Complex	Big Mints	Chew	Small Mints	Vitamin B	Candy		Magnesium	Calcium	B-Complex	Big Mints	Chew	Small Mints	Vitamin B	Candy	
	(a) Audio only $\varnothing$ 58.75 %									(b) Tactile only $\varnothing$ 51.25 %									(c) Multimodal $\varnothing$ 71.63 %								
Magnesium	0.42	0.21	0.02	0.19	0.03	0.01	0.05	0.07		0.49	0.07	0.13	0.09	0.05	0.14	0.0	0.01		0.7	0.08	0.09	0.07	0.04	0.0	0.01	0.01	0.02
Calcium	0.1	0.68	0.0	0.06	0.0	0.01	0.03	0.12		0.12	0.46	0.02	0.15	0.1	0.06	0.01	0.09		0.15	0.72	0.01	0.02	0.0	0.0	0.01	0.01	0.09
B-Complex	0.07	0.02	0.61	0.01	0.0	0.18	0.09	0.01		0.03	0.0	0.66	0.03	0.03	0.24	0.02	0.0		0.08	0.01	0.62	0.05	0.03	0.12	0.09	0.0	
Big Mints	0.11	0.05	0.1	0.5	0.05	0.03	0.05	0.11		0.02	0.05	0.03	0.44	0.08	0.23	0.09	0.08		0.05	0.06	0.04	0.56	0.12	0.05	0.09	0.04	
Chew	0.14	0.05	0.0	0.14	0.51	0.0	0.06	0.09		0.08	0.11	0.04	0.12	0.35	0.18	0.07	0.05		0.07	0.03	0.02	0.1	0.68	0.02	0.02	0.07	
Small Mints	0.01	0.02	0.07	0.02	0.01	0.83	0.02	0.03		0.05	0.02	0.12	0.07	0.06	0.63	0.03	0.02		0.0	0.01	0.02	0.01	0.01	0.92	0.03	0.01	
Vitamin B	0.08	0.09	0.05	0.06	0.0	0.07	0.47	0.18		0.0	0.0	0.05	0.05	0.05	0.19	0.63	0.01		0.0	0.0	0.04	0.08	0.06	0.01	0.74	0.08	
Candy	0.06	0.11	0.02	0.05	0.02	0.03	0.04	0.68		0.02	0.04	0.02	0.19	0.17	0.11	0.01	0.44		0.0	0.03	0.0	0.04	0.03	0.02	0.09	0.79	

## 7 Conclusion

We successfully combined the tactile information with noisy audio data to improve the accuracy compared to approaches regarding the individual signals. The effect of combining different modalities in a single network is distinctly visible in Tab. 2. The classification errors of the unimodal approaches could be reduced significantly by fusing the single modalities. While the advantage of multimodal classification over unimodal audio-based classification is insignificant without noise in the signal, even so, it could be an improvement for noisy environments. By considering tactile information, our work extends the approach by Eppe et al. [17] who classify materials in optically identical capsules just by considering audio data. Specifically, we show that taking tactile signals into account improves the robustness of neural network-based classification, even if the accuracy resulting from only the tactile modality is significantly lower than the accuracy of the audio network without noise. The neural network architecture used in this work was deliberately chosen to be simple as the main goal was to highlight the effect of multi sensor fusion of the audio and tactile modalities.

The overall classification accuracy of our approach without acoustic noise is slightly lower than the results provided by Eppe et al. [17] and Chen et al. [14]. However, the focus of our work was to show to what extent tactile data can improve the auditory classification performance under noisy conditions, and not to obtain a good absolute performance value. Furthermore, both approaches collected more data than we do and used more diverse materials.

Currently, the amount of pills in the container is not considered by the proposed approach. Potential future work involves the extension of our neural architecture to also estimate the filling state of the medicine containers. Furthermore, we will consider *interactive sensing* in the future. Specifically, we would like to investigate how we can actively modulate the shaking motion on-line during the classification process to further optimize the classification. A potential method to realize this is reinforcement learning [30,31].

## References

1. Nakamura, T., Nagai, T., Iwahashi, N.: Multimodal Object Categorization by a Robot. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. (2007) 2415–2420
2. Pieropan, A., Salvi, G., Pauwels, K., Kjellstrm, H.: Audio-Visual Classification and Detection of Human Manipulation Actions. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. (2014) 3045–3052
3. Sun, F., Liu, C., Huang, W., Zhang, J.: Object Classification and Grasp Planning using Visual and Tactile Sensing. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **46**(7) (2016) 969–979
4. Eppe, M., Kerzel, M., Griffiths, S., Ng, H.G., Wermter, S.: Combining Deep Learning for Visuo-motor Coordination with Object Detection and Tracking to Realize a High-level Interface for Robot Object-picking. In: IEEE RAS International Conference on Humanoid Robots (Humanoids). (2017) 612–617
5. Kerzel, M., Eppe, M., Heinrich, S., Abawi, F., Wermter, S.: Neurocognitive Shared Visuomotor Network for End-to-end Learning of Object Identification, Localization and Grasping on a Humanoid. In: IEEE Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). (2019) 19–24
6. Alfadhel, A., Khan, M.A., Cardoso de Freitas, S., Kosel, J.: Magnetic Tactile Sensor for Braille Reading. *IEEE Sensors Journal* **16**(24) (2016) 8700–8705
7. Litvak, D., Zigel, Y., Gannot, I.: Fall Detection of Elderly Through Floor Vibrations and Sound. In: 2008 30th annual international conference of the IEEE engineering in medicine and biology society, IEEE (2008) 4632–4635
8. Wettels, N., Santos, V.J., Johansson, R.S., Loeb, G.E.: Biomimetic Tactile Sensor Array. *Advanced Robotics* **22**(8) (2008) 829–849
9. Arian, M.S., Blaine, C.A., Loeb, G.E., Fishel, J.A.: Using the BioTac as a Tumor Localization Tool. In: 2014 IEEE Haptics Symposium (HAPTICS), IEEE (2014) 443–448
10. Su, Z., Hausman, K., Chebotar, Y., Molchanov, A., Loeb, G.E., Sukhatme, G.S., Schaal, S.: Force Estimation and Slip Detection/Classification for Grip Control using a Biomimetic Tactile Sensor. In: 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). (2015) 297–303
11. Fishel, J.A., Loeb, G.E.: Sensing Tactile Microvibrations with the BioTac Comparison with Human Sensitivity. In: 2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob). (2012) 1122–1127
12. Xu, D., Loeb, G.E., Fishel, J.A.: Tactile Identification of Objects using Bayesian Exploration. In: 2013 IEEE International Conference on Robotics and Automation, IEEE (2013) 3056–3061
13. Kerzel, M., Ali, M., Ng, H.G., Wermter, S.: Haptic Material Classification with a Multi-Channel Neural Network. In: International Joint Conference on Neural Networks (IJCNN). (2017) 439–446
14. Chen, C.L., Snyder, J.O., Ramadge, P.J.: Learning to Identify Container Contents through Tactile Vibration Signatures. In: 2016 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAN), IEEE (2016) 43–48
15. Durst, R.S., Krotkov, E.P.: Object Classification from Analysis of Impact Acoustics. In: Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots. Volume 1. (1995) 90–95

16. Luo, S., Zhu, L., Althoefer, K., Liu, H.: Knock-Knock: Acoustic Object Recognition by Using Stacked Denoising Autoencoders. *Neurocomputing* **267** (2017) 18–24
17. Eppe, M., Kerzel, M., Strahl, E., Wermter, S.: Deep Neural Object Analysis by Interactive Auditory Exploration with a Humanoid Robot. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE (2018) 284–289
18. Liang, H., Li, S., Ma, X., Hendrich, N., Gerkmann, T., Zhang, J.: Making Sense of Audio Vibration for Liquid Height Estimation in Robotic Pouring. arXiv preprint arXiv:1903.00650 (2019)
19. Sinapov, J., Schenck, C., Stoytchev, A.: Learning Relational Object Categories using Behavioral Exploration and Multimodal Perception. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), IEEE (2014) 5691–5698
20. Meeussen, W., Wise, M., Glaser, S., Chitta, S., McGann, C., Mihelich, P., Marder-Eppstein, E., Muja, M., Eruhimov, V., Foote, T., et al.: Autonomous Door Opening and Plugging in with a Personal Robot. In: 2010 IEEE International Conference on Robotics and Automation, IEEE (2010) 729–736
21. The Shadow Robot Company: The Shadow Dexterous Hand <https://www.shadowrobot.com/products/dexterous-hand/>. Last accessed 6 Oct 2020.
22. Wettels, N., Fishel, J.A., Loeb, G.E.: Multimodal Tactile Sensor. In: *The Human Hand as an Inspiration for Robot Hand Development*. Springer (2014) 405–429
23. Davis, S., Mermelstein, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE transactions on acoustics, speech, and signal processing* **28**(4) (1980) 357–366
24. Eppe, M., Alpay, T., Wermter, S.: Towards End-to-End Raw Audio Music Synthesis. In: *International Conference on Artificial Neural Networks (ICANN)*. (2018) 137–146
25. Strahl, E., Kerzel, M., Eppe, M., Griffiths, S.: Hear the Egg - Demonstrating Robotic Interactive Auditory Perception. In: *International Conference on Intelligent Robots and Systems (IROS)*. (2018) 5041
26. Bergstra, J., Yamins, D., Cox, D.D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. (2013)
27. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. *Neural computation* **9**(8) (1997) 1735–1780
28. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078 (2014)
29. Kingma, D.P., Ba, J.L.: Adam: a Method for Stochastic Optimization. In: *International Conference on Learning Representations (ICLR)*. (2015)
30. Eppe, M., Magg, S., Wermter, S.: Curriculum Goal Masking for Continuous Deep Reinforcement Learning. In: *International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. (2019) 183–188
31. Eppe, M., Nguyen, P.D.H., Wermter, S.: From Semantics to Execution: Integrating Action Planning with Reinforcement Learning for Robotic Causal Problem-solving. *Frontiers in Robotics and AI* **6** (2019)