Program Digest, 2019 IEEE International Ultrasonics Symposium (IUS)
Glasgow, Scotland, October 6-9, 2019

Neural Network-Based Detection of Ultrasonic Targets with Respect to Noise and Number of Sampling Positions

Patrick K. Kroh*, Ralph Simon† and Stefan J. Rupitsch*

* Chair of Sensor Technology, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
† Department of Ecological Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
patrick.k.kroh@fau.de, ralph.simon@vu.nl, stefan.rupitsch@fau.de

***Abstract*—A neural network-based approach for detection of sonar targets in air is presented in this contribution. Our approach may facilitate autonomous mobile systems to reliably detect and classify objects in their surrounding by using sonar information. This task might be extremely important in changing as well as unorganized environments. We perform target identification with long short-term memory networks as classifiers. Such are capable of dealing with variable numbers of echoes from multiple positions per input sequence, which facilitates more flexible operation. The impact of the number of recording positions per sequence and of noise is investigated. Furthermore, we demonstrate the improvement in classification performance in comparison to previously obtained results from multi-layer-perceptrons.**

***Index Terms*—sonar measurements; sonar detection; neural networks; feature extraction**

I. Introduction

Autonomous mobile vehicles – such as driverless transport systems and robots – are usually equipped with cameras and laser scanners to recognize their environment and navigate in it. Since the mentioned sensors are susceptible to variations in environment lighting conditions and optical properties, the systems are also often complemented with sonar sensors to improve navigation reliability. The sonar sensors normally only return distance measurements to the closest obstacles, but there is more information contained in the echoes, which may be exploited to aid navigation by ultrasonic scene as well as target recognition [1], [2]. For proper function of such approaches in real-world scenarios, a challenge that has to be resolved is that echo signals can be considerably affected by environment conditions, such as local changes in temperature, moisture, humidity, and air flow. To deal with such influences, we aim to utilize a machine learning approach with artificial neural networks, because these are known to cope well with noisy, distorted, and quantized input data. First results were obtained with multi-layer perceptron networks by the authors in [3]. In the currently presented work, the suitability of long short-term memory networks (LSTMs) is evaluated as these can deal with sequential input data of differing length [4] and should consequently be better suited for a flexible embedded implementation on a robot. For this purpose, the influence of input sequence data length as well as noise is analyzed.

II. Materials and Methods

A. Measurement Setup and Procedure

Data sets for classification were generated from echoes that were recorded with an automated laboratory setup (see Fig. 1). Thereby, many sequences of echoes could be recorded

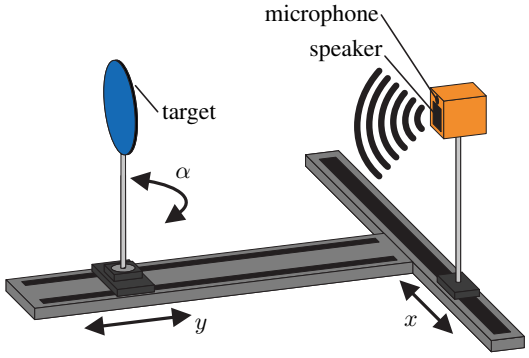

Fig. 1. Measurement setup with two translation stages (x and y) and a rotation stage (α). The targets are moved in y direction as well as rotated by the angle α. The speaker and the microphone are moved along the x axis. Thereby, echoes can be recorded at relative positions, which correspond to those of a robot passing by a target.

at various subsequent positions in order to emulate echoes received by a robot driving by a sonar target. For each sequence, a randomized start position within a distance of +/-0.5 m in x direction from the target, a randomized number of echoes per sequence, a random target angle (α within +/-60 deg), and random y target distance (within 0.5 m to 2.2 m) were set. The geometric x distance between subsequent echo recording positions of a sequence was set to be 5 cm. The translation stages had to stop at each recording position due to hardware limitations. This led to mechanical oscillations of the pillar on which the microphone as well as the speaker were mounted and, accordingly, further randomness in the data. For signal generation, recording, and processing, the following hardware was utilized: data acquisition device NI-USB 6356 (National

Instruments, $16\,\mathrm{bit}$, $1.25\,\mathrm{MSa\,s^{-1}}$), a desktop computer, an electrostatic Senscomp series 7000 ultrasonic speaker, and a $1/4''$ Bruel&Kjaer measurement microphone. The echo recordings were saved on a computer for processing and neural network training.

Rectified linear downchirp signals ($52\,\mathrm{kHz}$ down to $48\,\mathrm{kHz}$, $1\,\mathrm{ms}$) were emitted for pulse-echo operation. The sound pressure level was $110\,\mathrm{dB}$ re $20\,\mu\mathrm{Pa}$ at a distance of $1\,\mathrm{m}$ from the speaker.

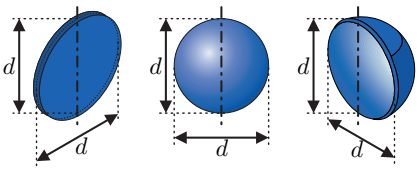

Fig. 2. Three different target shapes are used for classification: *a)* flat (discs), *b)* convex (spheres), and *c)* concave (hollow hemispheres). Classification was only performed with respect to shape and shall be independent of target size. As a consequence, each target shape was given in three different sizes with a characteristic dimension d of $60\,\mathrm{mm}$, $80\,\mathrm{mm}$, and $100\,\mathrm{mm}$ to make the classifiers generalize regarding target shape [5]. The hemispheres' hollow side is directed towards the speaker and microphone.

Flat (discs), convex (spheres), and concave (hollow hemispheres) shapes were chosen as target primitives (see Fig. 2), which are basic shapes into which arbitrary targets may be categorized for classification. The primitives' echo reflection behavior should differ considerably. At discs, primarily mirror-like reflection occurs, at spheres' surfaces, the impinging acoustic waves are reflected across a large angle due to the surface curvature, and at hemispheres, directed reflection mainly back into the impinging waves' direction occurs ("retroreflection"). More detailed explanations as well as analyses are given in [3] and [6]. Different from [3], here we use spheres instead of cylinders because tilt movements' influences on echoes are negligible for spheres. Also, greater (+/-$0.5\,\mathrm{m}$ instead of +/-$0.15\,\mathrm{m}$ in x direction) and randomized distances from the targets are investigated, which leads to more diverse, less redundant echo data. Details not mentioned here are identical to the ones from [3].

B. Data Set Generation

The echo recordings are processed and features are extracted to create samples for the training ($70\,\%$), validation ($15\,\%$) and test data sets ($15\,\%$), as shown in Fig. 3. In the current context, the term "feature" denotes any representation of data that may contain relevant information for classification. A more compact data representation leads to more efficient computation, since fewer parameters need to be learned and calculated. However, care must be taken because meaningful information will be discarded if a too compact representation is chosen. All features are combined in so-called feature vectors, which are the actual input to a classifier. Here, the term *sample* or *data set sample* denotes a set of associated feature vectors for an entire sequence of echoes and should not be confused with single *ADC samples*, which represent the output from an analog to digital converter (ADC).

The main preprocessing steps for a feature vector x_t are depicted as well as explained in Fig. 4. The specific feature selection is motivated as well as elaborated in detail in [3]. Spectrogram calculations are performed by means of a short-time fourier transform (STFT) with a window length of 256 ADC samples and $50\,\%$ overlap. Moreover, a threshold is defined for $\hat{r}_{yx}$, which has to be reached for at least one echo signal in a sequence. Otherwise, the corresponding data set sample is discarded. For the currently examined data, a suitable threshold was found to be $12\,\mathrm{dB}$ above the noise RMS level in the pulse-compressed echoes.

Data sets were generated with constant echo numbers per sequence (from one to ten) as well as with constant additional white noise levels for sequences of five echo positions (from $50\,\mathrm{dB}$ to $100\,\mathrm{dB}$ re $20\,\mu\mathrm{Pa}$). Furthermore, we generated example data sets to gain an impression of general neural network performance for miscellaneous parameters:

- "EXnb": variable number of echoes per sequence (from one to ten), randomized additional noise with a peak noise level of $70\,\mathrm{dB}$ re $20\,\mu\mathrm{Pa}$;
- "EXnbss": same as *EXnb* except for five times subsampling, which results in a sampling rate of $250\,\mathrm{kSa\,s^{-1}}$ and an accordingly set STFT window length of 32 ADC samples;
- "EXwb": same as *EXnb*, but for wideband downchirp excitation signals ($100\,\mathrm{kHz}$ to $52\,\mathrm{kHz}$)
- "EXnbMLP": same as *EXnb*, but with a constant number of five echoes per sequence, so that multi-layer perceptrons (MLPs) can be trained for comparison.

Each data set comprises 4500 samples, which are evenly distributed among the available sonar targets (500 data set samples per sonar target). Data set sizes decreased for large noise levels (above $80\,\mathrm{dB}$ re $20\,\mu\mathrm{Pa}$) to about 3000 samples because the noise threshold for target discovery increased.

C. Classification with Artificial Neural Networks

LSTMs are deployed as shape classifiers, whose main asset is that sequences with variable length can be used as input data. This is a major advantage in comparison to simple feed-forward networks, whose input feature vectors' dimensions must all be fixed. With an LSTM, it is even possible to generate classification output that is updated for each subsequent echo. As a consequence, well identifiable sonar targets can be detected quite fast and less easily identifiable targets may be detected after more echo recordings, which could then further improve navigation performance. The basic structure of an LSTM block is depicted as well as explained in Fig. 5.

The LSTM networks that were evaluated in the presented work were selected with a randomized parameter study. The best performing networks comprise 32 hidden units and were trained with a stochastic gradient descent optimization function with momentum. For training, the learning rate is 0.01, the momentum is 0.9, the mini batch size is 512, and training was executed for 4000 epochs. Among the evaluated networks, the shallow ones without additional hidden layers showed best results. For these networks, h_t is directly fed into an

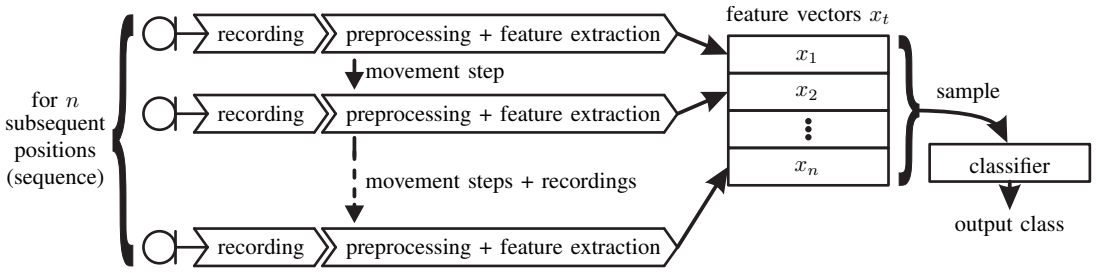

Fig. 3. Classification procedure; from echo recording and sample generation to classification itself.

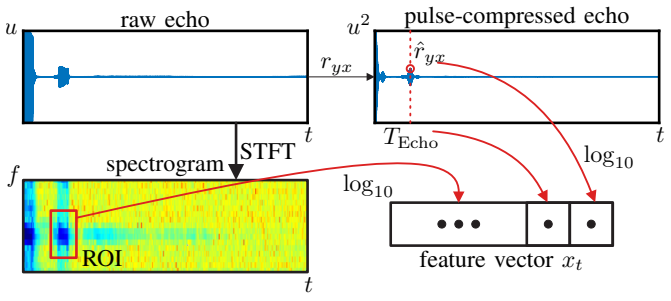

Fig. 4. Preprocessing and feature extraction; from raw echo (input voltage u) to feature vector. First, pulse-compression is performed by cross-correlation (r_{yx}) with a previously recorded excitation signal. The peak ($\hat{r}_{yx}$) that belongs to the sonar target's echo is selected from the pulse-compressed echo and its time delay is set as propagation delay assumption T_{Echo}. In addition, a spectrogram of the raw signal is calculated with a short-time fourier transform (STFT). A region of Interest (ROI) is selected, which is centered around T_{Echo}. ROIs have got a duration of 2 ms and cover the excitation frequency range. The resulting feature vector x_t comprises a concatenation of the logarithmized, flattened ROI, T_{Echo} as well as the logarithmized peak value $\hat{r}_{yx}$.

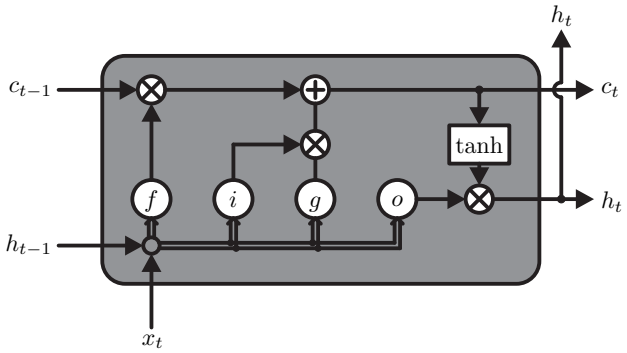

Fig. 5. Structure of a long short-term memory network (LSTM) block. Memory is realized by hidden state variables (c_t: cell state, h_t: hidden/output state) and gates (f, i, g, o), which can enable modification of the variables depending on variables from the previous state (c_{t-1}, h_{t-1}) as well as learned activation; f: forget gate with sigmoid activation, i: input gate with sigmoid activation, g: cell candidate gate with $\tanh$ activation, o: output gate with sigmoid activation. h_t may then be input to succeeding networks or network layers, such as multi-layer perceptrons (MLPs) or convolution, fully connected, and classification layers. Essentially, each gates itself is a fully connected perceptron layer [7].

output layer for classification (fully connected layer, softmax function, and classification layer). Work is currently going on to find suitable deep structures, which may yield increased performance. MLP parameters are identical to the ones from [3]. For each parameter configuration, five neural networks were trained on four different data sets each, adding up to 20 trained networks per evaluated parameter configuration.

III. Results and Discussion

We evaluated the neural network performance by total accuracy and AUC (area under curve for receiver operating characteristic ROC) for the test sets. The last classification outputs returned by the LSTMs for each sequence were used as labels. While the total accuracy is a measure of a classifier's total performance, AUC is a measure of individual class performance. Both measures are to be maximized towards 100 %. Total accuracy is the ratio between the number of correctly classified samples and the total number of samples. AUC is more complex and a thorough explanation is outside the scope of this paper. Please see [8] for details.

It can be seen from Fig. 6 that accuracy increases for larger numbers of echoes per sequence but that also a single echo contains much relevant information. It is also obvious that the employed LSTMs' performance clearly surpasses the MLPs'. An important factor here is the number of learned parameters, which scales linearly with the number of echoes per sequence for MLPs but is constant for LSTMs (compare also l in Table I). Another observation is that hollow hemispheres show the best results and seem to be clearly distinguishable from spheres as well as discs (Fig. 7).

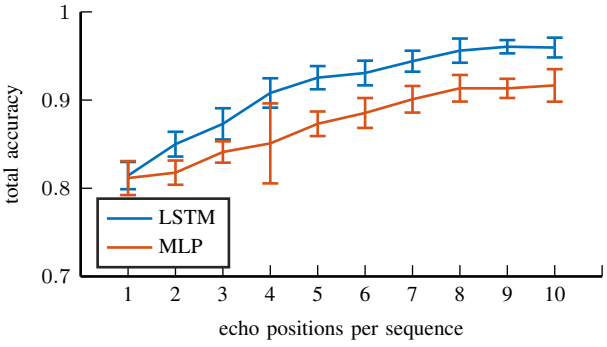

Fig. 6. Total network accuracy with respect to the number of positions per echo sequence for shape classification; comparison of LSTMs and MLPs; mean and standard deviation based on 20 trained networks for each data point.

The influence of additional random noise can bee seen in Fig. 8 and Fig. 9. It appears that the decrease of accuracy up to 80 dB is mainly caused by ambiguity between discs

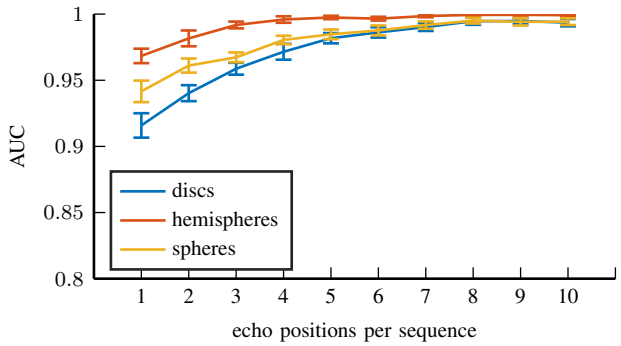

Fig. 7. Area under curve (AUC) of receiver operating characteristic (ROC) for shape classes with respect to the number of positions per echo sequence; for the LSTM networks from Fig. 6.

and spheres. This is in contrast to hemispheres, which can be identified quite well even for high noise levels (Fig. 9). Hemispheres' retroreflecting properties may play an important role here, since these lead to better echo signal-to-noise ratios. The noise range was chosen to start at 50 dB because a noise level without added noise was observed between 45 dB and 55 dB.

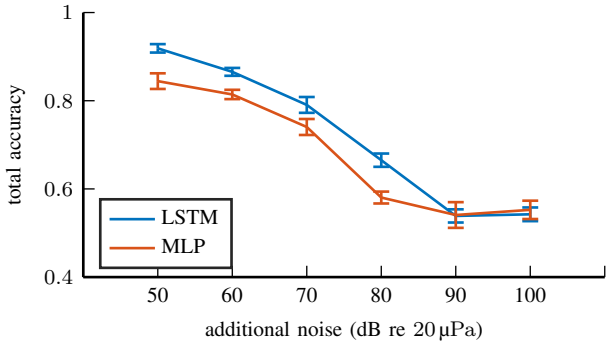

Fig. 8. Total network accuracy with respect to additional white noise for shape classification; for LSTMs and MLPs; five echoes per sequence; mean and standard deviation based on 20 trained networks for each data point.

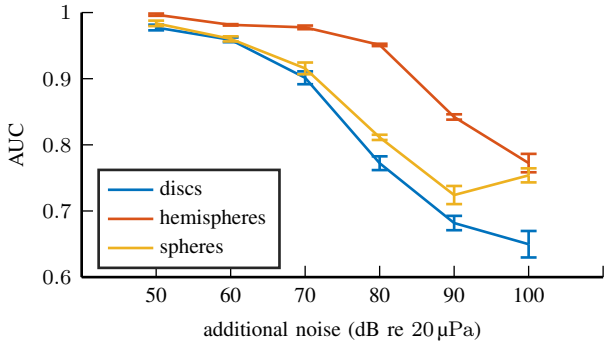

Fig. 9. AUC of ROC for different shape classes with respect to additional white noise; for the LSTM networks from Fig. 8.

TABLE I
PERFORMANCE MEASURES FOR EXAMPLE DATA SETS

	EXnb	EXnbss	EXwb	EXnbMLP
Accuracy mean	83.19 %	73.37 %	67.54 %	78.47 %
Accuracy std	1.25 %	1.73 %	7.15 %	6.13 %
'di' AUC mean	93.10 %	85.62 %	77.73 %	87.73 %
'di' AUC std	0.76 %	1.35 %	7.10 %	6.74 %
'he' AUC mean	98.32 %	95.43 %	95.87 %	97.73 %
'he' AUC std	0.34 %	0.66 %	1.60 %	0.69 %
'sp' AUC mean	94.19 %	88.22 %	82.91 %	88.59 %
'sp' AUC std	0.88 %	1.28 %	7.53 %	6.28 %
l	94	107	457	470

'di': discs. 'he': hemispheres. 'sp': spheres.
l: Feature vector length; $x_t \in \mathbb{R}^l$.

From Table I, one can see that the current implementation of EXnb provides the best overall values and that hemispheres can be well identified for all parameter configurations. We assume that better results for wideband excitation signals may be achieved than with the current LSTM implementation (EXwb), as is indicated by the results from [3].

IV. Conclusion

Successful use of LSTMs for sonar target classification in air was shown and an increase in performance as well as flexibility in contrast to previously evaluated MLPs could be demonstrated. Analyses of echo recording numbers as well as noise demonstrate that concave targets can be identified already with few echoes and even a considerable amount of noise. Such targets may thus be well suited as passive additional artificial acoustic navigation points. Classification of flat and convex targets is also possible, but requires more echo recordings and is more susceptible to noise. The results regarding noise performance motivate an embedded implementation of a sonar system on a mobile robot with commercial off-the-shelf components, such as piezoelectric transducers and MEMS microphones, which is currently under development.

References

[1] J. Steckel and H. Peremans, "Batslam: Simultaneous localization and mapping using biomimetic sonar," *PloS one*, vol. 8, no. 1, p. e54076, 2013.
[2] I. Eliakim, Z. Cohen, G. Kosa, and Y. Yovel, "A fully autonomous terrestrial bat-like acoustic robot," *PLoS computational biology*, vol. 14, no. 9, p. e1006406, 2018.
[3] P. K. Kroh, R. Simon, and S. J. Rupitsch, "Classification of sonar targets in air: A neural network approach," *Sensors (Basel, Switzerland)*, vol. 19, no. 5, 2019.
[4] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. [Online]. Available: http://arxiv.org/pdf/1404.7828v4
[5] D. von Helversen, "Object classification by echolocation in nectar feeding bats: Size-independent generalization of shape," *Journal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiology*, vol. 190, no. 7, pp. 515–521, 2004.
[6] R. Simon, M. W. Holderied, and O. von Helversen, "Size discrimination of hollow hemispheres by echolocation in a nectar feeding bat," *The Journal of experimental biology*, vol. 209, no. Pt 18, pp. 3599–3609, 2006.
[7] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2017.
[8] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.