Segmentation of Radar-Recorded Heart Sound Signals Using Bidirectional LSTM Networks

Kilin Shi[1]*, Sven Schellenberger[2], Leon Weber[3], Jan Philipp Wiedemann[1], Fabian Michler[1],
Tobias Steigleder[4], Anke Malessa[4], Fabian Lurz[1], Christoph Ostgathe[4], Robert Weigel[1],
and Alexander Koelpin[2]

Abstract— **Sounds caused by the action of the heart reflect both its health as well as deficiencies and are examined by physicians since antiquity. Pathologies of the valves, e.g. insufficiencies and stenosis, cardiac effusion, arrhythmia, inflammation of the surrounding tissue and other diagnosis can be reached by experienced physicians. However, practice is needed to assess the findings correctly. Furthermore, stethoscopes do not allow for long-term monitoring of a patient. Recently, radar technology has shown the ability to perform continuous touchless and thereby burden-free heart sound measurements. In order to perform automated classification of the signals, the first and most important step is to segment the heart sounds into their physiological phases. This paper examines the use of different Long Short-Term Memory (LSTM) architectures for this purpose based on a large dataset of radar-recorded heart sounds gathered from 30 different test persons in a clinical study. The best-performing network, a bidirectional LSTM, achieves a sample-wise accuracy of 93.4 % and a F1 score for the first heart sound of 95.8 %.**

I. Introduction

When thinking of doctors, the stethoscope immediately comes to mind as a characteristic feature. Physicians use it to listen to the sounds of a patient's heart, which may signify the action of a healthy heart, named heart sounds, as well as various cardiac pathologies, called heart murmurs. Heart sounds occur at every single heartbeat. In healthy persons, there are two physiological heart sounds: the first (S1) and the second (S2) heart sound. Both heart sounds occur at fixed points in time during the cardiac cycle and are caused by specific physiological phenomena of the heart's action. S1 is caused by the contraction of the heart's muscle at the end of the diastole when the ventricle is completely filled. In temporal comparison with the ECG the S1 starts with the R-peak. S2 is caused by closure of the semilunar valves of the aorta and truncus pulmonalis at the end of the expulsive phase and occurs simultaneously with the end of the T-wave in the ECG.

Pathological changes in the heart or the heart valves lead to heart murmurs. When a doctor listens to the heart, he or she checks for these pathological sounds. However, the validity of the assessment strongly depends on the experience of the physician. An objective and automated classification algorithm would not have this disadvantage since it would have access to a comprehensive database which is constantly expanded. Furthermore, these checkups only allow for a short-term analysis. In order to perform long-term recordings, different devices are needed. Recently, it has been shown that radar systems can record heart sounds without the need of permanent contact [1]. Using this technology, long-term recordings as well as an objective and automated analysis can be performed. This can be realized by a computer-aided technology which basically consists of two steps: (a) segmentation of the recorded sounds of the heart and (b) subsequent classification into normal / abnormal sounds of the heart using features that are calculated based on the segmentation. To ensure correct classification, an accurate segmentation procedure is crucial and is considered a challenging task due to noise artifacts or physiological abnormalities. During this step, heart sounds are segmented into the fundamental four temporal phases "S1" → "Systole" → "S2" → "Diastole".

Although a lot of attempts at segmentation algorithms of heart sounds have already been done, most of them require a priori knowledge, e.g., about the average length of the single segments [2]–[4]. This paper introduces an algorithm, which is capable of heart sound segmentation without any predefined parameters by using a bidirectional Long Short-Term Memory (LSTM) network. Although attempts with recurrent neuronal networks have also been made [5], there has been no optimization of the hyperparameters. Furthermore, this is the first attempt to work on radar-recorded heart sounds instead of heart sound signals recorded with a digital stethoscope. This is a crucial difference since radar systems measure distance information while the stethoscope measurand is proportional to the acceleration.

The research project GUARDIAN is supported by the Federal Ministry of Education and Research, Berlin, Germany, project grant No. 16SV7694.

The study was approved by the ethics committee of the Friedrich-Alexander-Universität Erlangen-Nürnberg (No. 85_15B). All research was performed in accordance with relevant guidelines and regulations. The informed consent was obtained from all subjects in human trials.

[1] Kilin Shi, Jan Philipp Wiedemann, Robert Weigel and Fabian Lurz are with the Institute for Electronic Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen 91058, Germany.

[2] Sven Schellenberger and Alexander Koelpin are with the Chair for Electronics and Sensors Systems, Brandenburg University of Technology, Cottbus 03046, Germany.

[3] Leon Weber is with the Computer Science Department, Humboldt-Universität zu Berlin, Berlin 10099, Germany.

[4] T. Steigleder, A. Malessa, C. Ostgathe are with the Palliative Care Department, Universitätsklinikum Erlangen, Erlangen 91054, Germany

*Corresponding author e-mail: kilin.shi@fau.de.

II. Long Short-Term Memory

An LSTM is a recurrent neural network which, in principle, can be used to model non-linear long-term dependencies between elements of a discrete time series [6]. The network produces an embedding $h_t \in \mathbb{R}^{d'}$ for each input $x_t \in \mathbb{R}^d$. At each time step t the network receives two inputs x_t, h_{t-1}. x_t are the features derived from the input signal at t, while h_{t-1} is the output of the LSTM computed in the preceding step. Both inputs are used to determine the states of the input gate i_t, the forget gate f_t and the output gate o_t, as well as the cell c_t. The cell is responsible for retaining some representation of the preceding elements of the time series $x_1, x_2, \dots, x_{t-1}$. The input gate modulates the amount to which the current input influences the state of the cell, while the forget gate may gradually reset the cell's state. The output h_t is derived from the current state of the cell and the output gate o_t. The exact formulae for all components are as follows [7]:

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) && (1)\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) && (2)\\
c_t &= (1 - i_t) \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + \\
&\qquad W_{hc}h_{t-1} + b_c) && (3)\\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) && (4)\\
h_t &= o_t \odot tanh(c_t) && (5)
\end{aligned}$$

where all Ws and bs are trainable parameters, σ denotes the element-wise sigmoid function and $\odot$ is the element-wise product. A visualization of the computation graph is shown in Fig. 1.

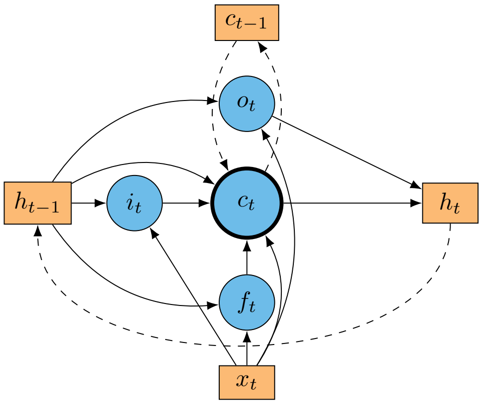


Fig. 1. Computation flow of an LSTM network. The gates o_t, i_t, f_t and the cell c_t are depicted as circles, while the input and output values x_t, h_t, h_{t-1} and c_{t-1} of the computation step are drawn as rectangles. Dashed lines imply a recurrence relation. [8]

When computing the embedding at t a standard LSTM can only take preceding inputs $x_1, x_2, \dots, x_t$ into account, but the subsequent inputs x_{t+1}, x_T might also carry relevant information. Thus, a bidirectional LSTM (biLSTM) [9] is also utilized which contains two LSTM layers each computing an embedding for the forward sequence $x_1, x_2, \dots, x_t$ and the backward sequence $x_T, x_{T-1}, \dots, x_t$. The final embedding for x_t is obtained by the concatenation of both respective embeddings for x_t.

III. Model and Training

Essentially, the proposed model is a two-layer neural network. The first layer is the biLSTM with 200 units for each direction which produces an embedding for each of the time steps. The resulting embedding h_t is fed into a fully-connected layer with four classes followed by a softmax layer which yields a vector of the predicted scores for each class:

$$\hat{y}_i = \frac{\exp(w_i \cdot h_t + b_i)}{\sum_{k=1}^{4} \exp(w_k \cdot h_t + b_k)}. \tag{6}$$

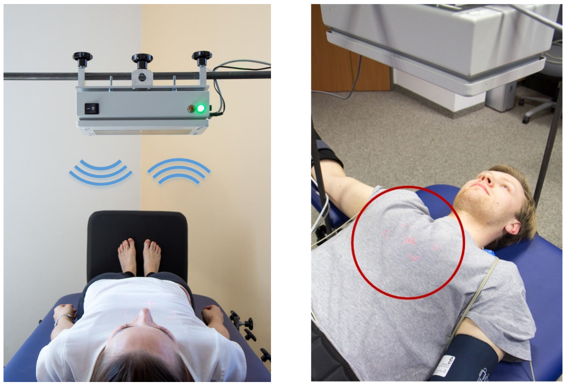
As a loss function l, standard multi-class cross-entropy is employed:

$$l(y, \hat{y}) = \sum_{k=1}^{4} y_k \log \hat{y}_k. \tag{7}$$

This objective is optimized for 100 epochs with vanilla stochastic gradient descent using an initial learning rate of 0.01. 100 epochs are chosen conservatively since no more substantial changes are noticed during training after around 40 to 50 epochs. No early stopping criteria is defined. If the validation loss does not decrease for 20 epochs, the learning rate is decreased by a factor of 10. The norm of the gradients is clipped to a maximum of 1 in order to avoid numerical instability. In order to combat overfitting, dropout [10] is used between the biLSTM and the fully-connected layer with a drop probability of 0.2.

IV. Data Acquisition and Methodology

Overall, 30 healthy test persons were measured for approximately 10 minutes each. The average age of all test persons is 30.7 years with a standard deviation of 9.9 years. The test persons consist of 16 female and 14 male persons. The average body mass index is 23.3 kg/m^2 with a standard deviation of 3.3 kg/m^2. All measurements were performed in the university hospital under medical supervision.



(a) (b)

Fig. 2. Pictures of the measurement setup. (a) A bistatic radar is placed above the lying test persons at a distance of approximately 40 cm to 50 cm from the antennas to the chest surface. (b) To facilitate the positioning of the antenna on the thorax, a red laser pattern is projected onto the surface.

The measurement scenario can be seen in Fig. 2. The radar system is placed above the lying test persons at a fixed distance. A red laser pattern is projected onto the thorax to visualize the focus spot of the antennas.

All heart sound signals were recorded using a Six-Port-based radar architecture [11]. It is realized as a bistatic radar system with one receive (RX) and one transmit (TX) antenna is used. An oscillator generates a high-frequency signal at 24 GHz which is split into two parts. One part is sent to the TX antenna, reflected at the skin surface of the thorax, received by the RX antenna and fed into the Six-Port receiver. The other, smaller portion is directly coupled into the Six-Port as a reference signal. Both the reference signal as well as the received signal are then superimposed under four relative and static phase shifts of 0°, 90°, 180° and 270°. These four output signals are down-converted into baseband using diode power detectors. The resulting signals $B_{3\,...\,6}$ form two pairs of differential signals which are orthogonal to each other. They can be interpreted as in-phase and quadrature components I and Q of a complex number $\underline{Z}$. To retrieve the relative distance change of the object, the argument of $\underline{Z}$, which represents the relative phase shift $\Delta\sigma$, has to be calculated [11]:

$$\Delta\sigma = \arg\{\underline{Z}\} = \arg\{(B_5 - B_6) + j(B_3 - B_4)\} \tag{8}$$

$$= \arctan(\frac{B_3 - B_4}{B_5 - B_6}). \tag{9}$$

Using $\Delta\sigma$ and the known wavelength λ, the relative distance change Δx can be calculated by:

$$\Delta x = \frac{\Delta\sigma}{2\pi} \cdot \frac{\lambda}{2}. \tag{10}$$

The raw I and Q data from the radar are directly transferred to a computer and all post-processing is done in MATLAB. After performing the above-mentioned arctangent demodulation, the raw distance signal is filtered in a range of 10 Hz to 80 Hz using a 4th order butterworth filter to retrieve the heart sound component from the radar signal [1]. In order to perform heart sound segmentation using the LSTM networks, features have to be calculated from the raw heart sound signal. For this purpose, three features are extracted: the homomorphic envelogram (HoEnv), the Hilbert envelope (HiEnv) and the power spectral density envelope (PSDEnv). The HoEnv has been already used by other segmentation algorithms and constitutes the exponentiated low-pass filtered natural logarithm of the heart sound signal [12], [13]. The HiEnv is calculated from the absolute value of the Hilbert transformation [14]. As last feature, the PSDEnv is derived from calculating the mean PSD between 40 Hz to 60 Hz in overlapping windows of 0.05 s length and 50 % overlap. The frequencies are derived from the fact that the majority of the frequency content is around 50 Hz [15].

V. Results

For the purpose of training and testing the LSTM network, the data is split into segments with a length of 10 s. Overall, the database consists of 1890 such segments.

Scores are calculated using three-fold cross-validation: the complete dataset is randomly split into three parts, in each iteration, two parts (1260 samples) are used for training and one part (630 samples) for testing. After each set has been used for testing, the final scores are calculated on all test sets.

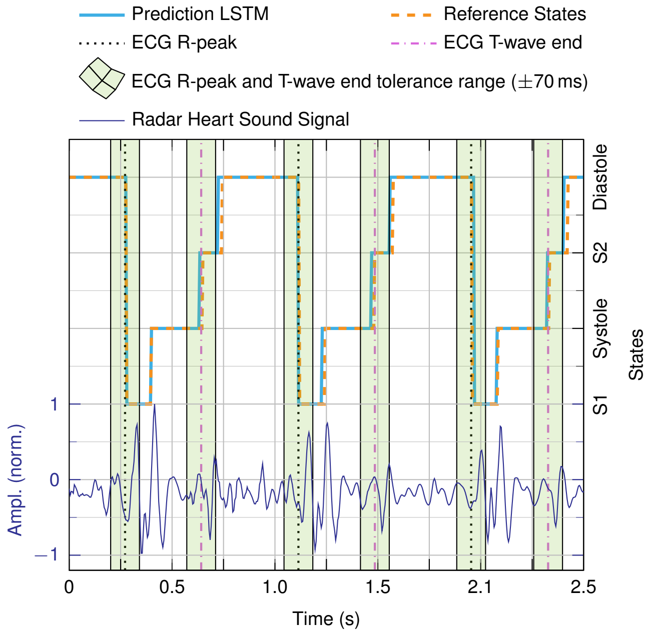


Fig. 3. Segmentation result using the biLSTM algorithm. The plot shows the raw radar heart sound signal, the predicted states and the reference states. The shaded areas indicate the tolerance ranges around the ECG R-peaks and T-wave ends in which a heart sound is counted as true positive.

To illustrate the scoring process, one exemplary segmentation on the test set using the biLSTM algorithm is shown in Fig. 3. Plotted are the heart sound signal, the reference ECG points in time and the reference states which are calculated using these points as well as the predicted states of the biLSTM algorithm. R-peaks and T-wave ends are found using the algorithm proposed by Zhang *et al.* [16]. Each heart sound corresponds to one reference point of the ECG. While the first heart sound starts right after the ECG R-Peak, the second heart sound starts immediately after the end of the T-wave. Therefore, the deviation between the predicted start of S1 and the R-peak as well as the deviation between the predicted start of S2 and the T-wave end is taken as key figure. A tolerance range of 70 ms is chosen for a heart sound to be correctly detected as true positive (TP). This is within the recognized tolerance range for ECG R-peak detection of 150 ms [17] and is further shortened to 70 ms. If no heart sound is found inside this range, a false negative (FN) is counted. The same applies the other way around: if no reference heart sound is found within this range around a predicted one, a false positive (FP) is counted. In this example, all S1 and S2 are found correctly.

Scores such as the overall F1 score, the F1 scores for the first and second heart sound separately (F1 S1 and F1 S2) as well as sensitivity, precision and accuracy are calculated using the mean scores of all single segments. Furthermore, by comparing the reference states and the predicted states, a sample-wise accuracy can be calculated. In addition, the 95 % confidence intervals (CI) of the scores are indicated. Specificity is not determined since true negatives are not

TABLE I

Mean scores including the 95 % CI for the test sets using different configurations and cross-validation

Configuration	**F1**	**F1 S1**	**F1 S2**	**Sensitivity**	**Precision**	**Accuracy**	**Accuracy (sw)**[a]
LSTM with 200 hidden units	84.0±0.7%	85.6±0.8%	82.6±1.0%	88.0±0.6%	81.0±0.8%	75.0±0.9%	88.5±0.3%
biLSTM with 50 hidden units	88.1±0.6%	95.4±0.4%	81.1±1.0%	90.0±0.6%	86.4±0.7%	81.1±0.9%	91.7±0.3%
biLSTM with 100 hidden units	87.8±0.6%	95.4±0.4%	80.7±1.0%	90.3±0.6%	85.8±0.7%	80.7±0.9%	92.3±0.3%
biLSTM with 200 hidden units	87.7±0.6%	95.8±0.4%	80.1±1.0%	90.6±0.6%	85.3±0.7%	80.5±0.9%	93.4±0.3%
biLSTM with 300 hidden units	87.2±0.6%	95.3±0.4%	79.6±1.0%	90.3±0.6%	84.6±0.7%	79.7±0.9%	92.6±0.3%
biLSTM with 200 hidden units and Hyperparameter Variation[b]	88.1±0.6%	94.2±0.5%	82.3±0.9%	90.2±0.6%	86.4±0.7%	81.0±0.8%	90.1±0.3%

[a] Sample-wise accuracy. [b] Gradient threshold 5, drop learn rate after 5 epochs, drop learn rate by a factor of 5.

determinable. The reference states are calculated with the help of the ECG signal as described in [3].

Different hyperparameter configurations shall be evaluated. For this purpose, different setups are tested. First of all, a standard LSTM layer is used instead of the biLSTM layer. Next, the number of hidden units is varied between 100, 200 and 300. Furthermore, a different set of hyperparameters for learning is chosen: The maximum gradient threshold is set to 5 instead of 1, the learning rate is decreased by a factor of 5 (instead of 10) after 5 (instead of 20) epochs if validation loss does not decrease. The resulting scores are shown in Table I. As it can be seen, the biLSTM networks perform better than the single LSTM network. Only for the F1 S2 score, the single LSTM network can reach higher scores. However, the sample-wise accuracy as well as the F1 S1 are considered as the most important scores since the first one gives an impression on the overall performance while the second one is important if the heart rate or any related values are to be calculated. Under these aspects, the biLSTM network with 200 hidden units and the learning hyperparameters configured as described in Section III is indeed the best-performing one. For this configuration, an overall F1 score of 87.7 % for both heart sounds and a F1 S1 score of 95.8 % is achieved. Furthermore, the biLSTM reaches a sample-wise accuracy of 93.4 %.

VI. Conclusion

Heart sounds are an important physiological signal whose analysis provides important information about the state of health of a person's heart. Radar technology enables touch-free and continuous monitoring of the heart sound signals. Subsequently, an automated analysis and classification of these signals allow for an objective assessment. The first and most important step is a correct segmentation of the four phases that occur during each cardiac cycle. This paper evaluated different LSTM architectures for this task as LSTMs have the advantage that no a priori information is needed. All parameters are learned during the training step. In summary, the biLSTM implementation with 200 hidden units reached the highest scores. As a next step, this segmentation algorithm might be combined with a subsequent classification algorithm. For this task, it would be required to also gather data from persons with heart diseases such as stenoses or insufficiencies.

References

[1] C. Will, K. Shi, S. Schellenberger, T. Steigleder, F. Michler, J. Fuchs, R. Weigel, C. Ostgathe, and A. Koelpin, "Radar-based heart sound detection," *Sci. Rep.*, vol. 8, no. 1, p. 11551, 2018.

[2] S. E. Schmidt, E. Toft, C. Holst-Hansen, C. Graff, and J. J. Struijk, "Segmentation of heart sound recordings from an electronic stethoscope by a duration dependent hidden-markov model," in *2008 35th Proc. Comput. Cardiol.*, Sept. 2008, pp. 345–348.

[3] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 822–832, Apr. 2016.

[4] C. D. Papadaniil and L. J. Hadjileontiadis, "Efficient heart sound segmentation and extraction using ensemble empirical mode decomposition and kurtosis features," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1138–1152, July 2014.

[5] E. Messner, M. Zhrer, and F. Pernkopf, "Heart sound segmentation - an event detection approach using deep recurrent neural networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1964–1974, Sept. 2018.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[7] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[8] K. Greff, R. K. Srivastava, J. Koutnk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[9] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5-6, pp. 602–610, 2005.

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[11] A. Koelpin, F. Lurz, S. Linz, S. Mann, C. Will, and S. Lindner, "Six-Port based interferometry for precise radar and sensing applications," *Sensors*, vol. 16, no. 10, p. 1556, 2016.

[12] C. N. Gupta, R. Palaniappan, S. Swaminathan, and S. M. Krishnan, "Neural network classification of homomorphic segmented heart sounds," *Appl. Soft Comput.*, vol. 7, no. 1, pp. 286–297, 2007.

[13] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in *2005 32th Proc. Comput. Cardiol.*, Sept. 2005, pp. 957–960.

[14] D. Kumar, P. Carvalho, M. Antunes, R. Paiva, and J. Henriques, "Noise detection during heart sound recording using periodicity signatures," *Physiol. Meas.*, vol. 32, no. 5, p. 599, 2011.

[15] P. Arnott, G. Pfeiffer, and M. Tavel, "Spectral analysis of heart sounds: relationships between some physical characteristics and frequency spectra of first and second heart sounds in normals and hypertensives," *J. Biomed. Eng.*, vol. 6, no. 2, pp. 121–128, 1984.

[16] Q. Zhang, A. I. Manriquez, C. Medigue, Y. Papelier, and M. Sorine, "An algorithm for robust and efficient location of t-wave ends in electrocardiograms," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2544–2552, Dec. 2006.

[17] ANSI/AAMI, *Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms*, Std. EC57, 2012.