Deep Unsupervised Representation Learning for Abnormal Heart Sound Classification

Shahin Amiriparian[1,2], Maximilian Schmitt[1], Nicholas Cummins[1], Kun Qian[1,2], Fengquan Dong[4], Björn Schuller[1,3]

shahin.amiriparian@tum.de

***Abstract*— Given the world-wide prevalence of heart disease, the robust and automatic detection of abnormal heart sounds could have profound effects on patient care and outcomes. In this regard, a comparison of conventional and state-of-the-art deep learning based computer audition paradigms for the audio classification task of normal, mild abnormalities, and moderate/severe abnormalities as present in phonocardiogram recordings, is presented herein. In particular, we explore the suitability of deep feature representations as learnt by sequence to sequence autoencoders based on the AUDEEP toolkit. Key results, gained on the new Heart Sounds Shenzhen corpus, indicate that a fused combination of deep unsupervised features is well suited to the three-way classification problem, achieving our highest unweighted average recall of 47.9 % on the test partition.**

***Index Terms*— Deep learning, unsupervised feature representation, abnormal heart sound detection**

I. INTRODUCTION

The human body produces a myriad of acoustic sounds that directly reflect changes in our physiological and pathological states and traits. For example, the phonocardiogram (PCG), arguably one of the earliest technical approaches towards analysing bio-signals, is the most fundamental method for diagnosing a variety of cardiovascular disorders, such as coronary heart disease, arrhythmia, and hypertension [1].

Computational audio understanding (computer audition) techniques have the potential to produce supporting technologies for cardiologists and general practitioners to help increase the clinical efficacy of auscultation, thus helping to reduce the high societal burden associated with heart diseases [2]. Moreover, advances in mobile and wearable recording and sensing devices are increasing the reliability and feasibility of remote diagnostic and monitor solutions [3].

Herein, we investigate if state-of-the-art computer audition paradigms can be applied to classify heart sounds. In particular, we explore the suitability of deep unsupervised feature representation learning; to the best of the authors' knowledge, this is the first time such a study has been undertaken. We compare two non-deep approaches, a conventional acoustic feature set [4] and a Bag-of-Audio-Words (BOAW) approach [5], with the AUDEEP toolkit which uses recurrent sequence to sequence autoencoders to learn deep unsupervised feature representations from raw audio [6], [7].

Our approaches are verified on the *Heart Sounds Shenzhen* (HSS) corpus, a novel database of 422.82 minutes of heart sound recordings collected from 170 participants (cf. Section III). The database allows for the three-way classification of heart sounds, namely, normal, mild abnormalities, and moderate/severe abnormalities.

The rest of this paper is organised as follows: a brief overview of related works is given in Section II, and the new HSS corpus is introduced in Section III. Our proposed recurrent sequence to sequence autoencoder is outlined in Section IV. The experimental settings and results are then given in Section V. Finally, the conclusions and outline of our future research plans are presented in Section VI.

II. RELATION TO PRIOR WORK

Our baseline feature representations, the brute-force knowledge driven COMPARE feature set [8], and the data driven BOAW representations have been used in a range of audio-health detection systems, for instance, snore sound classification [9]. Further, the combination of the COMPARE feature set and a Support Vector Machine (SVM) classifier is widely used in the field of computational paralinguistics.

As in most fields of computer audition, deep learning based solutions are starting to have a major impact in terms of achievable results as evidenced by entrants in the 2016 Computing in Cardiology Challenge; see [10] for a recent overview. To date, these approaches have been limited to the use of both Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) based classification systems [11]–[13]. However, deep learning can be used for *unsupervised feature generation*, in which meaningful and task specific features are automatically generated [14].

Sequential data, such as audio, often poses challenges for deep representation learning, as the underlying networks typically require fixed dimensionality inputs. The representation learning solution presented herein addresses this issue through the use of sequence to sequence autoencoders based on learning with RNNs. The advantages of such an approach have been shown in tasks like acoustic scene classification

[1]Shahin Amiriparian, Maximilian Schmitt, Nicholas Cummins, and Björn Schuller are with the ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany {shahin.amiriparian, maximilian.schmitt, nicholas.cummins, bjoern.schuller, kun.qian}@informatik.uni-augsburg.de

[2]Shahin Amiriparian and Kun Qian are also with the Machine Intelligence & Signal Processing Group, Technische Universität München, Germany

[3]Björn Schuller is also with GLAM – Group on Language, Audio and Music, Imperial College London, U. K.

[4]Fengquan Dong is with the Shenzhen University General Hospital, Shenzhen, P. R. China.

TABLE I
CLASS DISTRIBUTION PER PARTITION.

Partition	normal	mild	moderate/severe	SUM
Train.	84	276	142	502
Devel.	32	98	50	180
Test	-	-	-	163

and sound event classification [6], [7], but have yet to be verified for abnormal heart sound detection.

III. HEART SOUND DATASET

Our results are based on the HSS corpus which has recently been made available through the Interspeech 2018 Computational Paralinguistics Challenge[1]. The HSS corpus contains 845 recordings (with 30 seconds on average) representing 422.82 minutes. The heart recordings were collected using the electronic stethoscope from one of four locations: (i) the auscultatory mitral area, (ii) the aortic valve auscultation area, (iii) the pulmonary valve auscultation area, and (iv) the auscultatory area of the tricuspid valve. The recordings were collected from 170 independent subjects (55 female and 115 male), from mostly older individuals (ages range from 21 to 88 with the mean age being 65.44 years, and standard deviation of 13.24 years) with varying health conditions, including coronary heart disease, heart failure, arrhythmia, hypertension, hyperthyroid, and valvular heart disease.

As outlined, the corpus has been divided into three classes: (i) normal, (ii) mild, and (iii) moderate/severe, as diagnosed by specialists in heart diseases. These classes are divided into participant-independent training, development, and test sets with 502, 180, and 163 audio instances, respectively. The gender and age classes are evenly distributed. In summary, there are 100 normal, 35 mild, and 35 moderate/severe subjects (cf. Table I). As indicated, at the time of writing, the data was an active Interspeech Computational Paralinguistics Challenge dataset; therefore the divisions for the test partitions were not publicly available.

IV. DEEP UNSUPERVISED REPRESENTATION LEARNING

A high-level structure of our deep unsupervised representation learning approach is given in Figure 1. First, Mel-spectrograms are obtained from the raw heartbeat recordings (cf. Figure 1a). A sequence to sequence autoencoder is then trained on these extracted spectra (cf. Figure 1b) that are considered as time-dependent sequences of frequency vectors. After autoencoder training, the learnt representations of the Mel-spectrograms are then generated for use as feature vectors for the corresponding instances (cf. Figure 1c). Finally, we train a classifier (cf. Figure 1d) on the feature sets to predict the labels of the heartbeat recordings.

A. Spectrogram Extraction

First, we generate the power spectra of heartbeat recordings using periodic Hann windows with variable width w and overlap $0.5\,w$. Subsequently, a given number N_{mel} of log-scaled Mel frequency bands are computed from the spectra. Mel-spectra features have previously been shown to be effective for heart sound classification [12]. Finally, we normalise the Mel-spectra values in $[-1;1]$, as the outputs of the autoencoder are constrained to this interval.

Furthermore, important acoustic cues related to the class label may be obscured by background noise during the recording of the heartbeats. Hence, we investigate whether removing some background noise from the spectrograms improves system performance. This is achieved by clipping amplitudes below a certain threshold.

B. Recurrent Sequence to Sequence Autoencoders

Recurrent sequence to sequence autoencoders are applied for unsupervised representation learning of the extracted Mel-spectra [6], [7], [15]. We consider the Mel-spectra as time-dependant sequences of frequency vectors in $[-1;1]^{N_{mel}}$. Each sequence represent the amplitudes of the N_{mel} Mel frequency bands within one audio segment and is then fed to a multilayered *encoder* RNN. The hidden state of the encoder is then updated regarding to the input frequency vector. Accordingly, the final hidden state of the encoder RNN comprises information regarding the full input sequence. This final hidden state is reconstructed applying a fully connected layer. Another multilayered *decoder* RNN is used to rebuild the original input sequence from the reconstructed feature (cf. Figure 2). For full details, the interested reader is referred to [6].

The encoder RNN has N_{layer} layers and each layer contains N_{unit} Gated Recurrent Units (GRUs). During training, we apply the root mean square error (RMSE) between the target sequence as the objective function and the decoder output. In order to cope with overfitting, we apply dropout [16] to the inputs and outputs of the recurrent layers, but not to the hidden states. After the training process, we extract the activations of the fully connected layer as the learnt representations of the Mel-spectra.

V. EXPERIMENTAL SETTINGS AND RESULTS

Unless otherwise stated, all experimental settings are outlined in the following subsections.

A. Recurrent Sequence to Sequence Autoencoders

We implemented the approach described in Section IV using the AUDEEP toolkit[2]. The toolkit is written in Python, and depends on TENSORFLOW[3] for the core autoencoder implementations. We use the Adam optimiser to train the autoencoders with a fixed learning rate of 0.001 [17] for 64 epochs in batches of 256 samples. A dropout of 20 % has been applied to the outputs of each recurrent layer. Moreover, gradients with absolute value above 2 are clipped [15].

[1] http://emotion-research.net/sigs/speech-sig/is18-compare

[2] https://github.com/auDeep/auDeep

[3] https://www.tensorflow.org/

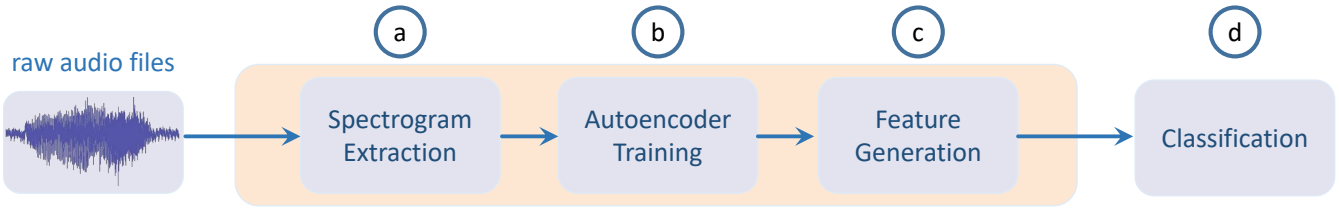

Fig. 1. Structure of our deep representation learning and classification system with recurrent autoencoders. The approach is – except for the final classification – fully unsupervised. The procedure is described in details in Section IV.

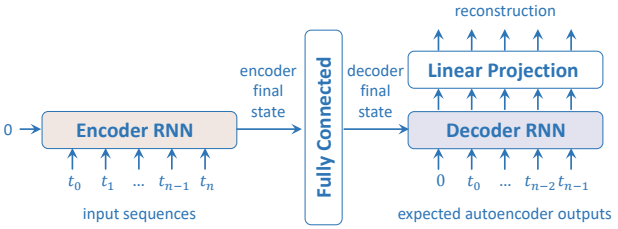

Fig. 2. A high-level structure of the applied recurrent autoencoder.

Our deep learning system contains a wide range of adjustable hyperparameters that prohibits an exhaustive analysis of the parameter space. Accordingly, we choose suitable values for the hyperparameters in various stages, using the findings of our initial experiments to bootstrap the process.

First, we selected a suitable autoencoder configuration with the ideal number of recurrent layers N_{layer}, the number of GRUs per layer N_{unit}, and either bidirectional or unidirectional decoder and encoder RNNs. The sequence to sequence autoencoders are trained on Mel-spectrograms generated with the window width $w = 0.32$ seconds, the window overlap $0.5\,w = 0.16$ seconds, and $N_{mel} = 128$ Mel frequency bands, with amplitude clipping below a certain threshold. We applied thresholds under -30 dB, -45 dB, -60 dB, and -75 dB. During our preliminary evaluations, these chosen parameters provided reasonable results. We exhaustively evaluated $N_{layer} \in \{2,3,4\}$, $N_{unit} \in \{64,128,256,512\}$ and all combinations of bidirectional or unidirectional encoder and decoder RNNs. The highest Unweighted Average Recall (UAR) was achieved when using $N_{layer} = 2$ layers and $N_{unit} = 256$ units with a unidirectional encoder RNN and a bidirectional decoder RNN.

In the second development stage, we optimised the window width w applied for spectrogram creation. We use the autoencoder configuration specified in the first stage. We then evaluate the window width w between 0.08 and 0.36 seconds with a step size of 0.04 seconds. The window overlap is set to $0.5\,w$. During initial experimentation, we observed that the windows size $w = 0.32$ seconds, as set in the previous step, provided the strongest UAR. We speculate that using window sizes shorter than $w < 0.32$ seconds result in weaker representation due to the lack of discriminating information in the shorter audio segments. For larger values of $w > 0.32$ seconds, we observed that the classification accuracy dropped again. This could have been caused by the larger window width blurring the short-term dynamics of the heartbeat sounds.

In the final optimisation stage, we tested various numbers of Mel frequency bands $N_{mel} \in \{16,32,64,128,256\}$. With larger values of N_{mel} the classification accuracy rises until it stops increasing for $N_{mel} > 128$. For this reason, we choose $N_{mel} = 128$ to reduce the amount of data which the system has to process.

B. ComParE Acoustic Feature Set

Further results presented are based on the *Interspeech 2016 Computational Paralinguistics Challenge* feature set ComParE [4]. This feature set comprises a range of prosodic, spectral, cepstral, and voice quality *low-level descriptor* (LLD) – prosodic, spectral, cepstral, and voice quality – contours, to which statistical functionals such as the mean, standard deviation, percentiles and quartiles, linear regression descriptors, and local minima/maxima related descriptors are applied to produce a 6 373 dimensional static feature vector.

C. Bag-of-Audio-Words

Bag-of-Audio-Words (BoAW) computed using the toolkit openXBOW [5], are also tested. BoAW involves the quantisation of acoustic LLDs to form a sparse fixed length histogram (bag) representation of an audio clip. Due to the quantisation step, which can be considered a quasi lowpass filtering operation, BoAW representations are generally considered more robust than LLDs.

All BoAW representations were generated from the 65 LLDs and corresponding deltas in the ComParE feature set. Prior to quantisation the LLDs were normalised to zero mean and unit variance. All codebooks were learnt using openXBOW random sampling setting with *codebook size* (*cs*) 250, 500, and 1000 considered.

D. Classification Set-Up

In order to predict the class labels for the audio instances in the heartbeat corpus, we train a linear SVM classifier using the Sequential Minimal Optimisation (SMO) approach implemented in WEKA 3.8.2 [18]).

Features were scaled to zero mean and unit standard deviation, using the parameters from the training set. The complexity hyperparameter of the SVM was optimised in the range between 10^{-6} and 10^{-1} for our deep learning, ComParE, and BoAW approaches. The SVM complexity that performed the strongest on the development set was applied to train the final classifier with the fusion of the training and development sets. Due to small imbalances in the class distribution of our data (cf. Section III), all classification systems are evaluated using the *Unweighted Average Recall* (UAR) metric.

TABLE II

A COMPARISON OF ACCURACIES OF OUR SEQUENCE TO SEQUENCE AUTOENCODER SYSTEM WITH A COMPARE FEATURE SET AND A BOAW APPROACH. THE CHANCE LEVEL IS 33.3 % UAR.

System	Dimensionality	UAR [%] C	Devel.	Test
COMPARE	6 373	10^{-6}	41.1	44.8
		10^{-5}	44.5	45.6
		10^{-4}	50.3	46.4
		10^{-3}	44.5	40.4
		10^{-2}	43.2	41.7
BOAW	250	10^{-3}	43.1	43.4
	500	10^{-3}	42.3	47.2
	1000	10^{-2}	43.7	41.0
AUDEEP: Individual Feature Sets				
-30 dB	1 024	$2 \cdot 10^{-2}$	32.8	40.0
-45 dB	1 024	$5 \cdot 10^{-4}$	38.4	40.6
-60 dB	1 024	$6 \cdot 10^{-2}$	39.6	45.2
-75 dB	1 024	$8 \cdot 10^{-3}$	36.9	41.7
Fused	4 096	$4 \cdot 10^{-3}$	35.2	**47.9**

E. Results and Discussion

The strongest development set UAR, 50.3 % (cf. Table II), was achieved using a system based on the COMPARE feature set and a SVM complexity of 10^{-4}. However, this system had a noticeable drop in performance on the HSS test partition indicating possible overfitting. For the conventional (non-deep) approaches, the strongest test set partition UAR, 47.2 % (cf. Table II) was achieved using a BOAW approach with a codebook size of 500, and a SVM complexity of 10^{-3}.

For our deep recurrent approach, we extracted four feature sets by amplitude clipping below thresholds of -30 dB, -45 dB, -60 dB, and -75 dB (cf. Section IV-A). These learnt representations achieved a weaker performance than the conventional feature sets on the development partition. This could be due to the small amount of data for training the autoencoder. When comparing with the conventional approaches on the HSS test set, the learnt representations achieve equivalent performance. Moreover, an early fusion of the four learnt deep feature vectors obtain the highest UAR, 47.9 % (cf. Table II) on the test set. This result indicates the promise of deep representation learning for abnormal heart sound classification.

VI. CONCLUSIONS AND FUTURE WORK

Technologies based on state-of-the-art computer audition systems have the potential to aid the diagnosis of cardiovascular disorders. In this regard, the presented results indicate the suitability of deep learning to learn meaningful representations from phonocardiogram (PCG) recordings. We showed that fusing all deep representations after amplitude clipping, it is possible to outperform the conventional acoustic features for the task of abnormal heart sound detection. In future work, we will use other PCG databases, such as the Computing in Cardiology corpus [1], to provide further training material for our deep learning approaches.

VII. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu).

REFERENCES

[1] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, pp. 2181–2213, 2016.

[2] D. Mozaffarian, E. J. Benjamin, A. Go, D. K. Arnett, M. J. Blaha *et al.*, "Executive summary: Heart disease and stroke statistics-2016 update: A report from the american heart association," *Circulation*, vol. 133, no. 4, pp. 447–454, 2016.

[3] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.

[4] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[5] M. Schmitt and B. Schuller, "openXBOW — Introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, 2017, 5 pages.

[6] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *DCASE2017*. Munich, Germany: IEEE, Nov. 2017, pp. 17–21.

[7] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *arXiv preprint arXiv:1712.04382*, 2017.

[8] B. Schuller, S. Steidl, A. Batliner *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[9] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski *et al.*, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proc. INTERSPEECH 2017*, ISCA. Stockholm, Sweden: ISCA, August 2017, pp. 3442–3446.

[10] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *2016 Computing in Cardiology Conference (CinC)*, Sept 2016, pp. 609–612.

[11] R. Banerjee, A. D. Choudhury, P. Deshpande, S. Bhattacharya, A. Pal, and K. M. Mandana, "A robust dataset-agnostic heart disease classifier from phonocardiogram," in *Proc. of EMBC 17*, July 2017, pp. 4582–4585.

[12] V. Maknickas and A. Maknickas, "Recognition of normalabnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients," *Physiological Measurement*, vol. 38, no. 8, p. 1671, 2017.

[13] C. Schölzel and A. Dominik, "Can electrocardiogram classification be applied to phonocardiogram data? –an analysis using recurrent neural networks," in *2016 Computing in Cardiology Conference (CinC)*, Sept 2016, pp. 581–584.

[14] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112.

[16] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014, 15 pages.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.