Noise Masking Recurrent Neural Network for Respiratory Sound Classification

Kirill Kochetov[1], Evgeny Putin[1], Maksim Balashov[1], and Anatoly Shalyto[1]

[1]Computer Technologies Lab, ITMO University,
49 Kronverksky Pr, 197101 St. Petersburg, Russia
{kskochetov,eoputin,balashov,shalyto}@corp.ifmo.ru

Abstract. In this paper we propose a novel architecture of recurrent neural network called noise masking recurrent neural network (NMRNN) for lung sound classification. The model jointly learns to extract only important respiratory-like frames without redundant noise and then by exploiting this information is trained to classify lung sounds into four categories: normal, containing wheezes, crackles and both wheezes and crackles. We also compare the performance of our model with merely machine learning based models. As a result, the NMRNN model reaches state-of-the-art performance on recently introduced publicly available respiratory sound database.

Keywords: Respiratory Sound Classification, Recurrent Neural Networks, Deep Learning

1 Introduction

In the last decades many machine learning (ML) approaches have been introduced to analyze respiratory cycle sounds including crackles, coughs, wheezes [1–6]. However almost all conventional ML models solely rely on hand-crafted features. Furthermore, highly complex preprocessing steps are required to make use of designed features [4–6]. Thus, merely ML-based models may not be robust to external/internal noises in lung sounds and may not generalize their performance across different softwares and measuring devices. However, to be used in clinics respiratory tracking systems have to reach high classification accuracy.

From that perspective deep learning (DL) models [7] have gained a lot of attention in the community. DL-based models primary rely on high abstract representation of data that are learned through the training of models. Due to this fact that DL models reach state-of-the-art performance on the range of tasks including image recognition [8], speech recognition [9], time series forecasting [10].

In this work we propose an architecture of RNN called NMRNN that is trained in end-to-end manner to simultaneously detect noise in respiratory cycles and to classify lung sounds into several categories of breath like normal, containing wheezes, crackles or both. In other words, our model itself decides what information and from what time points it should use to make effective

prediction of respiratory sounds. The crucial feature of the NMRNN is that it is trained without applying any hand preprocessing stages like slicing on individual respiratory cycles. Through extensive testing, the proposed model have reached state-of-the-art performance on recently published large open database of lung sound records [11].

The rest of the paper is organized as follows. In Section 2, we review several notable works in respiratory sounds classification using ML and DL based models. Detailed description of NMRNN is given in Section 3. Sections 4 and 5 presents experiments overview with our results and comparative study with solely ML-based models. Conclusion and future work is discussed in Section 6.

2 Related Work

Recently a comprehensive comparative study of applying different ML models to automatic wheeze detection was done in [4]. Authors used a lot of models including feed-forward neural network, random forest (RF), support vector machine (SVM) and trained them on two datasets: phonopneumogram samples and the Dubrovnik General Hospital (DGH) dataset. To reduce the influence of cardiovascular and muscular noise Yule-Walker filter was employed followed by STFT procedure. Then, two types of features were extracted from the lung sounds: MFCC (Mel-frequency cepstral coefficients) features and some statistical features. Authors reported that their best model statistical features got 93.62%, 91.77% accuracies on phonopneumograms and DGH datasets, accordingly. Meanwhile, based on MFCC features SVM model reached 99% accuracy on both datasets.

In [12] authors proposed to use hidden Markov models (HMM) coupled with Gaussian mixture models (GMM) for classification of respiratory sounds into four categories: normal, containing wheezes, crackles and both crackles and wheezes. The main idea behind employing HMM was that HMM is able to take into account frame position in a sequence which leads to better accuracy comparing to GMM. To tackle with noise in sound records spectral subtraction technique [13] was applied. MFCC extracted from the records were used as input features to the model. In addition to MFCC features obtained in range from 50 Hz to 2000 Hz, the first time derivatives of MFCCs were used to track feature dynamics and to decorrelate feature vectors resulting in feature set with size 30. As a result, the ensemble model of 28 HMMs with 5 states and 1 Gaussian per state achieved 0.495 and 0.396 scores on the cross-validation and second evaluation score respectively. In both experiments different patients were used for training and testing, so it was honest validation and we can compare these results with ours.

One of the most successful attempt of applying DL models to the field of respiratory sound classification was done in [14]. Authors used convolutional neural networks (CNN) to detect wheezes in lung sound records. Firstly, respiratory records were augmented by biasing sound sample in several time frames. Then, STFT features were computed followed by standard normalization. Lastly, ob-

tained normalized spectrograms of lung sounds were used to train 2D CNN. The final model received 99% accuracy and 0.96 AUC on the dataset.

3 Method

Recurrent neural networks are a class of artificial neural networks (ANNs), that capable to process temporal data, such as sound and text. RNNs can use their internal state (memory) and feedback to process sequences of inputs.

One of the most popular variants of RNN is LSTM (Long short-term memory) and GRU (gated recurrent unit) networks [15, 16]. They are a dominant approaches that shows grand performance on sequence-related tasks such as NLP (Natural Language Processing) [17] and speech recognition [18].

We use both LSTM and GRU units for our experiments. The main idea of proposed approach is to adapt RNNs, which designed for time-scale data and can consider all information from sequential frames of input signal.

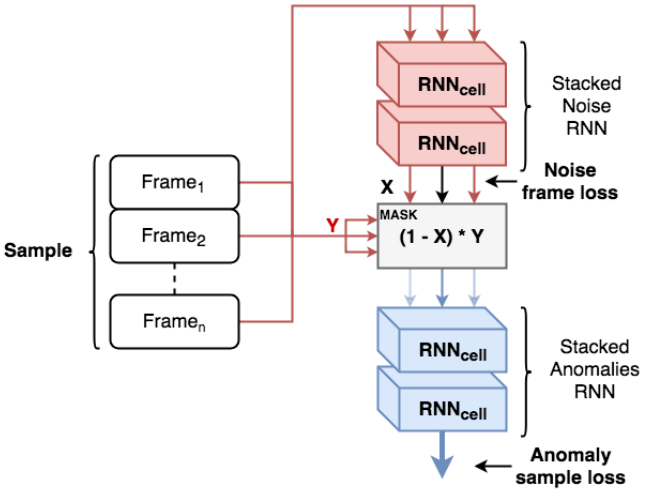

Fig. 1. Proposed approach. Stacked Noise RNN predicts one noise label per each frame using original MFCC data. MASK block makes attention on most important frames with respiratory cycles. Stacked Anomalies RNN predicts one anomaly label per each sample using highlighted data from MASK.

Our approach consists of three parts: noise classifier, respiratory (or anomalies) classifier and some kind of attention called MASK. Scheme of the proposed model is shown on figure 1.

First of all, before model training each sound sample was split on frames with equal length. There is only one anomaly label for sound sample and one noise label for each frame.

Noise classifier is a stacked RNN called NRNN, which predicts noise label for every frame from sample. NRNN optimize a cross-entropy loss calculated for each output.

$$L_{CE}(p,q) = -\sum p(x) \times log(q(x)) \tag{1}$$

Then predicted noise labels propagates through masking layer called MASK, where original frames multiplies with masking coefficient $(1-X) \times Y$, where X is predicted noise label ($X = 1$ for noise frame) and Y is a frame.

Anomaly classifier is a stacked RNN called ARNN, which predicts one anomaly label for one sample (all frames). ARNN takes highlighted frames from MASK block as input data and optimize a cross-entropy loss for one label per sample.

The final loss of the proposed architecture is following:

$$L_{model} = a_1 \times L_{CEnoise} + a_2 \times L_{CEanom} \tag{2}$$

Values of coefficients a_1 and a_2 based on idea, that the main goal of the model is anomaly classification, not noise classification, which only helps.

The proposed MASK mechanism is simple and efficient and was inspired by gating technique used in GRU cell, where memory needs to be rewritten on each time step using only important information from the input. NRNN paramenters were optimized using both NRNN and ARNN losses, so NRNN+MASK mechanism allows not only to mask noise frames, but to highlight useful subsamples with respiratory-like content. Attention mechanism used in current model is not the same as usually used for seq2seq models [19]. The main difference is that seq2seq attention mechanism commonly highlight important features from one latent vector, but our MASK rely on both predicted noise and anomaly labels. We conducted additional experiment to show that model with provided attention outperforms model without it in terms of classification metrics.

The crucial idea of proposed method is an ability to end-to-end classification without using any manual preprocessing steps like slicing respiratory on separate cycles. The only preprocessing step is splitting data on equal frames. Amount of frames does not matter too.

4 Experiments

In the study logistic regression (LR), random forest, gradient boosting machine (GBM) and SVM-based classifier [20] were used as baselines for comparison with proposed NMRNN.

4.1 Database

For training and evaluation the ICBHI Scientific Challenge database was used [11]. Database contains audio samples, collected independently by two research teams in two different countries over several years. The database consists of a total of 5.5 hours of recordings containing 6898 respiratory cycles, of which

1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes in 920 annotated audio samples from 126 subjects (patients). Some database summary presented in Table 1.

There are a lot of noise in sounds: 1840 noise cycles in all data and 1366 in AKGC417L data. It simulates real life conditions and made the classification algorithm more robust and stable for noise attack.

Table 1. Database summary. Recordings columns includes statistics about separate sound recordings data. Cycles columns includes statistics about individual respiratory cycles

Num of	**Recordings**		**Cycles**	
	All equipment	**AKGC417L**	**All equipment**	**AKGC417L**
Patients	126	56	126	56
Samples	920	683	6898	4697
Normal breath	287	196	3642	2226
Wheezes	134	77	886	512
Crackles	297	252	1864	1578
Wheezes & Crackles	202	158	506	381

4.2 Experiments setup

In this study we provided several experiments. Different data and preprocessing steps was used for them. The key idea of all experiments is to compare proposed approach with another machine learning models in different situations in terms of performance and robustness.

1. Simple noise binary classification experiment for initial model checking
2. 4-class anomalies classification using individual respiratory cycle as input
3. 4-class anomalies classification using sound samples with several respiratory cycles in each (end-to-end classification)

Goal of the first experiment is just to check NMRNN ability to learn respiratory and noise cycle intervals lengths and frequencies.

Goal of the second experiment is in comparison of our baseline models with recently proposed method [12]. Second experiment is good and clear, but it has one critical limitation: it is not end-to-end experiment, because first of all we need to split our sound on respiratory cycles, but there is no absolute universal solution for this task yet. So, for each new lung sound record we need to split it on respiratory cycles first.

For this reason third experiment was conducted. Goal of the this experiment is to check model ability to find what input information is important and where its located in big feature space. Model as end-to-end classifier needs to find respiratory-dependent features in the data by itself.

Also, there is two variations of data for each experiment. We use all data and data recorded only with AKGC417L microphone. The main idea of using second data type is to show, that models can achieve better performance using only one unbiased data source.

All experiments were conducted on a computer with Intel Core i7-6900 CPU with 128GB of RAM and NVIDIA GTX 1080Ti GPU.

4.3 Result evaluation

Due to the unbalanced data set, we sensitivity and specificity as statistical indicators of the effectiveness of the diagnostic test for the detection of sick and non-sick patients. Sensitivity, specificity and overall score were proposed in the original data set paper [11, 12].

Overall evaluation score can be formulated as:

$$Score = \frac{Sensitivity + Specificity}{2} \tag{3}$$

Cross-validation over patients was used to evaluate the results. The idea behind cross-validation is to divide the data set into disjoint training and validation subsets K in different but regular ways, after that a performance measure is evaluated as the mean value on all folds. Thus, results from cross-validation experiments are robust. We used 5-fold cross-validation and it is important to note that there is no patients from the train set in the test set on each split. So, we used honest real-oriented division of data for validation.

4.4 Preprocessing

To remove sounds caused by heartbeats, the signal components at low frequencies have to be suppressed. We use the high pass finite impulse response (FIR) filter with cutoff frequency $fc = 100$ Hz for remove sounds caused by heartbeat [12].

In this study MFCC was used as feature extractor. The lower and upper frequencies of processed content were cut to 50 and 2000 Hz respectively, because wheezes and crackles are in this interval [12]. Parameters frame length and frame step were both chosen equal to 0.05 second using grid search optimization.

Every sound sample from original database was sliced on pieces called frames with length of 0.5 seconds each. Every frame was split on 10 non-overlapping frames. Both frame length and frame step are 0.05 second. One MFCC set (13 values) was extracted from each frame. So, every piece is described by 130 MFCC features. Each frame and sample corresponds to a breathing (presence of anomaly) and noise label. There are four breathing classes in the database: normal breathing, breathing with wheezes, crackles and with both wheezes and crackles.

During anomaly classification using all frames (one label per sound) or subset of frames (one label per respiratory cycle) we want to predict existence of anomalies in the overall sound sample or in the only one respiratory cycle respectively. So, for baseline models each sound sample or respiratory cycle was reshaped

into a single flattened array. Taking into account different audio lengths, final data samples were cut or filled using standard padding technique and final. Also, augmentation technique (was proposed in [14]) with shifting was used for solving the problem of respiratory cycles localization. PCA (Principal Component Analysis) was used for dimensionality reduction (only for baseline models).

5 Results

For noise binary classification task NMRNN achieved 0.89 evaluation score compared with the best baseline model GBM, which achieved only 0.53 score. It can be explained by the ability of RNN to learn cycle and noise intervals length and frequency and use this information during prediction.

Table 2. Results of 4-class classification of each respiratory cycle. Metrics of Jakovljevic HMM was not provided with AKGC417L data

	All equipment			**AKGC417L**		
Model	**Sens**	**Spec**	**Score**	**Sens**	**Spec**	**Score**
GBM	0.476	0.554	0.515	0.534	0.568	0.551
LR	0.425	0.508	0.466	0.426	0.51	0.468
RF	0.438	0.538	0.488	0.483	0.521	0.502
SVM	0.49	0.502	0.496	0.502	0.518	0.51
Jakovljevic [12]	0.423	0.567	0.495	-	-	-
RNN (ours)	0.584	0.73	**0.657**	0.617	0.741	0.679

Results of 4-class classification of each respiratory cycle are presented in Table 2. There is a comparison of our baseline and NMRNN models with method proposed by Jakovljevi. All our models were trained on MFCC features. Performance of our models is similar with performance of Jakovljevic HMM [12], except for NMRNN, which outperforms competitors. So, it is correct to compare presented baseline models with proposed RNN-based approach in the next experiment. Also, models trained only on AKGC417L data show better scores as expected due to reduced bias of data distribution. It is important to note, that the second experiment is less complex than the third one, because of data manually sliced on respiratory cycles.

Results of end-to-end classification are provided in Table 3. NMRNN definitely outperform baseline methods in terms of presented metrics. The main reason is that RNN was designed to process such kind of data with temporal dependencies. Another models faces with problems of large dimensionality and localization of respiratory cycles. So, neither PCA and augmentation does not help to solve these problems, because baseline models are not adopted for unstable data with floating content such as sound with several respiratory cycles.

MASK block with noise classification increases performance on about 0.035 in terms of score. It can be explained by ability of final model to concentrate only

Table 3. Results of 4-class classification of each sound sample

	All equipment			AKGC417L		
Model	**Sens**	**Spec**	**Score**	**Sens**	**Spec**	**Score**
GBM	0.362	0.142	0.252	0.348	0.174	0.261
LR	0.348	0.184	0.266	0.366	0.236	0.301
RF	0.433	0.054	0.244	0.451	0.079	0.265
SVM	0.313	0.251	0.282	0.278	0.256	0.267
RNN (ours)	0.511	0.717	0.614	0.572	0.728	0.65
NMRNN (ours)	0.56	0.736	**0.648**	0.62	0.75	**0.685**

on frames with respiratory cycles, not with noise. Also, MASK block helps to distinguish false positive anomalies (biased noise) with real anomalies (crackles or wheezes) as justified on Figure 2.

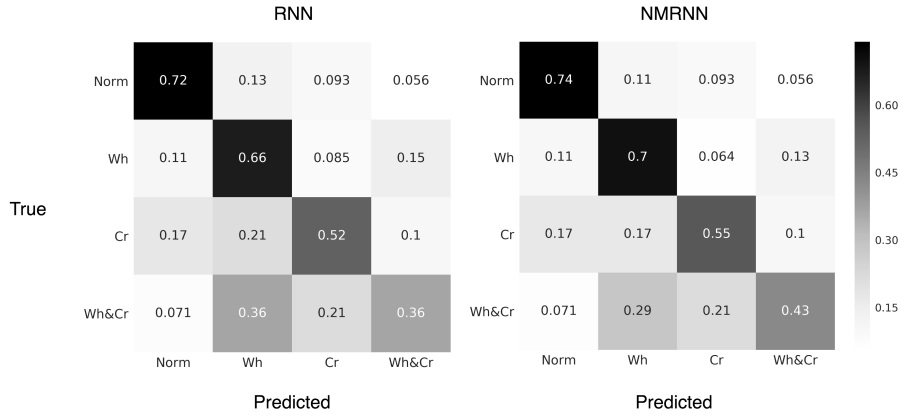

Fig. 2. Confusion matrices of RNN and NMRNN. MASK block helps to clarify some samples similarity by masking false positive anomalies detected in noise frames. Due to that both sensitivity and specificity was improved.

Models trained only on AKGC417L data have better performance as in the previous experiments. This suggests that model can be adopted for single source and can in theory boost performance with increasing of amount of unbiased data for training.

We used grid search [21] as optimization algorithm for finding best hyper-parameters for baseline and RNN-based models. So, the best RNN-based model with MASK block consists of 2-layer RNNs as both NRNN and ARNN parts with GRU cells with 256 units in each. Coefficients a_1 and a_2 from equation 2 are 0.3 and 0.7 respectively, which corresponds to the main task of the model (anomaly classification). Overall model architecture was trained using Adam [22] optimizer with $learning\_rate = 0.0001$.

6 Conclusion

In this paper we proposed RNN-based end-to-end model called NMRNN to detect different anomalies in lung sound data. The main contribution of this approach is that it is trained without applying any manual preprocessing steps using respiratory records of any lengths. NMRNN reaches state-of-the-art performance in comparison with another ML models on respiratory sound classification task and, including recently proposed [12], on individual respiratory cycle classification task.

Also, this study shows ability of model to learn cycle and noise intervals length and frequency. Experiments with AKGC417L microphone motivate to concentrate on single data source during creation of approach applicable in real life conditions.

MASK block is very powerful, so it allows to consider only relevant frames during classification. We assume, that the trained MASK mechanism is a superior direction of further improvement.

7 Acknowledgements

This work was financially supported by Goverment of Russian Federation (Grant 08-08)

References

1. M Bahoura and C Pelletier. Respiratory sounds classification using cepstral analysis and gaussian mixture models. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 1, pages 9–12. IEEE, 2004.
2. P Mayorga, C Druzgalski, RL Morelos, OH Gonzalez, and J Vidales. Acoustics based assessment of respiratory diseases using gmm classification. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 6312–6316. IEEE, 2010.
3. Rajkumar Palaniappan, Kenneth Sundaraj, and Sebastian Sundaraj. A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals. *BMC bioinformatics*, 15(1):223, 2014.
4. MARIO MILICEVIC, IGOR MAZIC, and MIRJANA BONKOVIC. Classification accuracy comparison of asthmatic wheezing sounds recorded under ideal and real-world conditions. In *15th International Conference on Artificial Intelligence, Knowledge Engineering and Databases (AIKED 2016), Venice*, 2016.
5. BM Rocha, L Mendes, I Chouvarda, P Carvalho, and RP Paiva. Detection of cough and adventitious respiratory sounds in audio recordings by internal sound analysis. In *Precision Medicine Powered by pHealth and Connected Health*, pages 51–55. Springer, 2018.
6. Gorkem Serbes, Sezer Ulukaya, and Yasemin P Kahya. An automated lung sound preprocessing and classification system based onspectral analysis methods. In *Precision Medicine Powered by pHealth and Connected Health*, pages 45–49. Springer, 2018.

7. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
8. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.
9. Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert. Analysis of cnn-based speech recognition system using raw speech as input. Technical report, Idiap, 2015.
10. Andreas S Weigend. *Time series prediction: forecasting the future and understanding the past*. Routledge, 2018.
11. BM Rocha, D Filos, L Mendes, I Vogiatzis, E Perantoni, E Kaimakamis, P Natsiavas, A Oliveira, C Jácome, A Marques, et al. A respiratory sound database for the development of automated classification. In *Precision Medicine Powered by pHealth and Connected Health*, pages 33–37. Springer, 2018.
12. N Jakovljević and T Lončar-Turukalo. Hidden markov model based respiratory sound classification. In *Precision Medicine Powered by pHealth and Connected Health*, pages 39–43. Springer, 2018.
13. Michael Berouti, Richard Schwartz, and John Makhoul. Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, volume 4, pages 208–211. IEEE, 1979.
14. Kirill Kochetov, Evgeny Putin, Svyatoslav Azizov, Ilya Skorobogatov, and Andrey Filchenkov. Wheeze detection using convolutional neural networks. In *Portuguese Conference on Artificial Intelligence*, pages 162–173. Springer, 2017.
15. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
16. Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
17. Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
18. Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.
19. Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.
20. T Hastie, R Tibshirani, and J Friedman. The elements of statistical learning 2nd edition, 2009.
21. James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.
22. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.