PhD Thesis

# A Holistic Approach to Multi-channel Lung Sound Classification

## Recording Hardware Development, Data Collection, and Deep Learning for Classification

conducted at the
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

in co-operation with
Division of Thoracic and Hyperbaric Surgery
Medical University of Graz, Austria

by
Dipl.-Ing. Elmar Messner, BSc

Supervisor:
Assoc.Prof. Dipl.-Ing. Dr.mont. Franz Pernkopf

Assessors/Examiners:
Assoc.Prof. Dipl.-Ing. Dr.mont. Franz Pernkopf
Asst.Prof. Miguel Tavares Coimbra, PhD

Graz, April 2, 2019

This work was supported by the government of Styria, Austria, under the project call HTI:Tech_for_Med and by the Austrian Science Fund (FWF) under the project number P27803-N15.

# Acknowledgement

I owe my deepest gratitude to my supervisor, Franz Pernkopf, who has supported me throughout my thesis with his guidance and knowledge. I greatly value his continuous and reliable support.

Many thanks go to our research cooperation partners from the Medical University of Graz: Dr. Paul Swatek, Dr. Melanie Fediuk, Dr. Stefan Scheidl, Prof. Freyja-Maria Smolle-Jüttner, and Prof. Horst Olschewski. I also thank the Clinical Trial Coordination Centre of the Medical University of Graz, especially Gabriele Pfaffenthaler and Dr. Astrid Friedel, for their support related to our clinical trial.

A special thanks goes to my brother Dr. Egon Messner for his medical advising, despite having no involvement in the research project. I really appreciate the numerous discussions.

I am grateful to Martin Hagmüller for his help especially in the initial stages of the research project and I thank Andreas Läßer for his assistance during hardware development.

Many thanks also go to the numerous people who volunteered as test subjects during the hardware development and/or as participants in our clinical trial.

I thank Wolfgang, Johannes, and all the other colleagues from the Signal Processing and Speech Communication Laboratory for providing such a good work atmosphere.

I also want to thank Prof. Miguel Coimbra for taking the long journey to Graz and being my examiner.

Last but not least, my sincere thanks go to my family and friends for supporting me in any and every possible way. Most importantly, I want to thank my mother Irma for her endless support.

# Abstract

Lung auscultation is an efficient and economic means to assess the state of the pulmonary organ. Pathological changes of the lung are tightly connected to characteristic sounds, often enabling fast and inexpensive diagnosis. To facilitate a more objective assessment of lung sounds for the diagnosis of pulmonary diseases/conditions, digital recording and post-processing techniques using computers have been a focus of intensive research over the past decades. In this thesis, we present a holistic approach to computer-aided lung sound analysis. In particular, we focus on multi-channel lung sound classification. We designed a lung sound recording device, conducted a clinical trial for data collection, and developed classification frameworks with a focus on deep neural networks.

We present a 16-channel recording device for airflow-aware lung sound recording. To this end, we developed a novel lung sound transducer and an appropriate attachment method realised as a foam pad. Compared to common approaches, we improved the usability and robustness against air- and body-borne noise. The device enables fast recording of lung sounds in noisy clinical environments without impeding daily routines.

Using the developed device, we conducted a clinical trial to record a multi-channel lung sound database. We included lung-healthy subjects and patients diagnosed with idiopathic pulmonary fibrosis.

For the classification of the recorded data, we consider two approaches. In our first approach, we present an event detection framework to detect adventitious sounds in lung sounds. In this context, due to a large public available heart sound dataset, we first introduce a new methodology for the segmentation of heart sounds. In particular, we propose an event detection approach with deep recurrent neural networks and show state-of-the-art performance carefully evaluated on the 2016 PhysioNet/CinC Challenge dataset. We then applied the final setup to adventitious sound and breathing phase detection in lung sounds. Our second approach for lung sound classification provides a direct diagnosis of the underlying disease. For this, we present a frame-wise classification framework to process full breathing cycles of multi-channel lung sound recordings with convolutional recurrent neural networks. The evaluation of both approaches on our multi-channel lung sound database shows promising results for the diagnosis of idiopathic pulmonary fibrosis.

# Kurzfassung

Die Lungenauskultation ist eine effiziente und wirtschaftliche Möglichkeit den Zustand des Lungenorgans zu beurteilen. Pathologische Veränderungen der Lunge sind eng mit charakteristischen Geräuschen verbunden, die oft eine schnelle und kostengünstige Diagnose erlauben. Um eine objektivere Beurteilung von Atemgeräuschen für die Diagnose pulmonaler Erkrankungen zu ermöglichen, waren computerunterstützte Aufzeichnungs- und Nachbearbeitungstechniken in den letzten Jahrzehnten Gegenstand intensiver Forschung. In der vorliegenden Arbeit wird ein ganzheitlicher Ansatz für die computerunterstützte Atemgeräuschanalyse, mit Fokus auf Klassifizierung von Mehrkanal-Aufnahmen, präsentiert. Zu diesem Zweck wurde zuerst ein entsprechendes Aufnahmegerät entwickelt, anschließend eine klinische Studie zur Datenerhebung durchgeführt, und darauf aufbauend Klassifikationssysteme unter Verwendung von künstlichen neuronalen Netzen entwickelt.

Das entwickelte Aufnahmegerät für Atemgeräusche ermöglicht die Erfassung von 16-Kanal-Aufnahmen über dem Rücken und eine simultane Atemflussaufzeichnung. Hierfür wurde ein neuartiger Sensor und eine geeignete Anbringungsmethode in Form eines Schaumstoffpolsters entwickelt. Im Vergleich zu herkömmlichen Ansätzen weist das Aufnahmegerät eine Verbesserung der Benutzerfreundlichkeit und der Robustheit gegenüber Körper- und Umgebungsgeräuschen auf. Das Gerät ermöglicht eine effiziente Erfassung von Atemgeräuschen, auch in lauten klinischen Umgebungen.

Unter Verwendung des entwickelten Aufnahmegeräts wurde eine klinische Studie zur Erstellung einer Mehrkanal-Atemgeräuschdatenbank durchgeführt. Diese Datenbank enthält Aufnahmen von lungengesunden Teilnehmern und von Patienten mit idiopathischer Lungenfibrose.

Wir präsentieren zwei Ansätze zur Klassifizierung von Atemgeräuschen. Der erste Ansatz beschäftigt sich mit der Detektion von pulmonalen Nebengeräuschen. Hierfür wurde zuerst eine neue Methodik für die Segmentierung von Herztönen entwickelt, wobei ein Ansatz zur Ereigniserkennung mit tiefen rekurrenten neuronalen Netzen vorgestellt wird. Experimentell werden State-of-the-Art-Ergebnisse auf dem PhysioNet/CinC Challenge-Datensatz von 2016 gezeigt. In einem zweiten Schritt wurde das Klassifikationssystem erfolgreich zur Ereigniserkennung von Nebengeräuschen und Atemphasen in Atemgeräuschen angewandt. Der zweite Ansatz zur Klassifizierung von Atemgeräuschen ist eine Methode zur direkten Diagnose der zugrunde liegenden Erkrankung. Das Klassifikationssystem verarbeitet vollständige Atemzyklen von Mehrkanal-Atemgeräuschen mit konvolutionellen rekurrenten neuronalen Netzen. Die Auswertung beider Klassifizierungsansätze auf der Mehrkanal-Atemgeräuschdatenbank zeigt vielversprechende Ergebnisse für die Diagnose von idiopathischer Lungenfibrose.

# Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

date

(signature)

# Contents

# 1

# Introduction

Lung auscultation is a noninvasive and fast means to assess the state of the pulmonary organ. However, traditional auscultation with a stethoscope has several disadvantages and limitations: Trained experts are rare, it is subjective (i.e. diagnosis depends on the experience of the physician), it cannot provide continuous monitoring, and a quiet environment is needed. Furthermore, the characteristics of the sounds are in the low frequency range, where the human hearing has limited sensitivity and is susceptible to noise artifacts [1]. Additionally, the intensity of the respiratory signal affects the masking of adventitious sounds, e.g. deep breaths may mask more crackles than superficial breaths [2]. Due to these facts, auscultation is simply used for a first screening of patients, since physicians have to rely on objective techniques enabling a reliable diagnosis, such as chest X-ray and computed tomography (CT) scan.

To facilitate a more *objective* assessment of lung sounds for the diagnosis of pulmonary diseases/conditions, digital recording and post-processing techniques using computers have been a focus of intensive research over the past decades [1–3]. Computational methods for the analysis of lung sounds have overcome many limitations of human auscultation and offer advantages for medical diagnosis, i.e. digital storage of lung sounds, monitoring of lung sounds in critical care settings or during surgery, computer-supported analysis, evaluation of long-term changes caused by certain diseases, and comparison among different sound recordings. Despite these advantages, computational lung sound analysis (CLSA) has not yet been used as a major tool for diagnosis of respiratory diseases [4].

Commercially available devices for lung sound recording are electronic stethoscopes. Compared to traditional stethoscopes, their main advantages are signal amplification and noise reduction to improve the listening experience [5]. They also allow for the sounds to be recorded and analysed on a computer afterwards. For this reason, electronic stethoscopes are widely used in lung sound research [6–8]. Other sensors include contact sensors or air-coupled microphones. The latter are either microphones with custom designed couplers or microphones inserted into stethoscope rubber tubes [9, 10]. Most research is performed on single-channel recordings, but some multi-channel approaches also exist [1]. The first (and formerly commercially available) device for multi-channel lung sound recording was the Stethographics STG 16 [11]. It enables 14-channel lung sound recording on the posterior chest, with two additional channels for the trachea and heart locations. Other multi-channel approaches found in the literature are a 14-channel device presented in [12] and a 25-channel device used in [13].

Even though lung sound research is a very active area, publicly available lung sound recordings are rare. According to [1], databases are either from online repositories [14, 15] or audio CDs companioning books [16–18]. The first large public database for the purpose of the development of classification algorithms was published in 2018 [19]. It contains 920 single-channel recordings from 126 subjects, with a total of 6898 respiration cycles. However, to the best of our knowledge, no publicly available multi-channel lung sound database exists.

Recently, research has been mainly focused on lung sound *classification* [1, 4]. The main task is the detection or classification of adventitious sounds. This can be performed on three different levels [1]: Detection and classification of adventitious sounds at a segment level (i.e. signal window segments are generated, features are extracted, and with random segments of adventitious and normal sounds classification is performed), classification at the event level (of manually isolated events of adventitious and normal lung sounds), and event detection at recording level. For this purpose, well established classifiers were widely applied, such as support vector machines (SVMs) [20, 21], Gaussian mixture models (GMMs) [22, 23], multilayer perceptrons (MLPs) [24, 25], random forests [26], k-nearest neighbours algorithm (k-NN) [27, 28], hidden Markov models (HMMs) [7, 8], and logistic regression [29]. Due to the recent success of deep learning in several audio applications, such as speech recognition [30] and acoustic event detection/scene classification [31], it also found its way to lung sound classification [32–34].

Several reviews related to CLSA have been published in recent years [1–4, 35, 36]. The state-of-the-art report in [2] concludes with the statement that it is necessary to find new markers to increase the efficiency of decision aid algorithms. Similarly, the authors in [3] identified the need for further research to promote its diagnostic utility in clinical settings. Most recently, a systematic review of 77 articles of automatic adventitious respiratory sound analysis was performed in [1]. The authors provide a list of challenges for future activities: in particular, CLSA requires adventitious sound monitoring as integral part for diagnosis, high accuracy algorithms (including feature extraction) for adventitious sound detection and classification, careful evaluation in real-life scenarios, and a portable easy-to-use device without the necessity of expert interaction. Furthermore, the review in [4] concluded that machine learning techniques are required to improve the accuracy and to promote the commercialisation as a product.

For a holistic view on lung sound classification, we identify several limitations. In the recording stage, we observe a lack of signal quality. This can be due to a low signal-to-noise ratio of typical sensors or the limitations of the rest of the recording hardware. Another reason is the superimposition of air- and body-borne noise, due to unsuitable recording environments and attachment methods. We also observe limitations regarding the measurement procedures. Single-channel recording with successive measurements to cover the whole pulmonary organ can be very time-consuming. Additionally, the breathing behaviour varies between successive recordings. This makes the assessment of differences in lung sound characteristics and intensity difficult. Also, airflow-unaware recording, i.e. ignoring the influence of airflow rate and breathing pattern on the signal characteristics, renders lung sound analysis even more challenging. Due to all these limitations, research regarding lung sound classification should highly benefit from a joint consideration of recording hardware, computational methods, and clinical evaluation issues.

## 1.1 Research Questions and Objectives

Traditional auscultation with a stethoscope is limited to a preliminary screening for lung diseases. However, our vision is to enable a reliable, easy-to-use lung sound analysis and decision support system for better assistance to patients and avoidance of diagnostic odysseys. It should enable a non-invasive, early detection of lung diseases and reduce the unwanted exposure to radiation due to X-rays or CT scans. Having this in mind, we pursue a holistic approach to multi-channel lung sound classification, with the simplified framework illustrated in Figure 1.1. Our main objectives and research questions are summarised as follows:

1. *Multi-channel Lung Sound Recording Device*: The aim is to develop a multi-channel lung sound recording device suitable to record a lung sound database within a clinical trial. Therefore, it should feature good usability and allow lung sounds to be recorded in real clinical environments, i.e. without the need for a specific recording room. Furthermore, it should fulfill the basic safety requirements for medical devices. The research questions include the development of a lung sound transducer and an appropriate attachment method. For multi-channel recording, we must find suitable recording positions and define the number of sensors needed. Furthermore, the influence of respiratory airflow on lung sounds has to be investigated.

2. *Lung Sound Database*: The aim is to record a multi-channel lung sound corpus for lung-healthy subjects and patients with selected lung diseases. This requires the design and organisation of a clinical trial. Hence, several design aspects have to be considered, such as suitability of subjects, the examination setting, and the measurement procedure itself.

3. *Lung Sound Classification Frameworks*: The aim is to develop classification frameworks for lung sounds with a focus on deep neural networks. Two approaches should be considered: (i) Event detection: A method to accurately detect acoustic events in lung sound recordings, such as adventitious lung sounds and breathing phase events (i.e. inspiration/expiration). (ii) Direct diagnosis of the underlying disease: A method to process multi-channel lung sound recordings that provides a diagnosis as output.

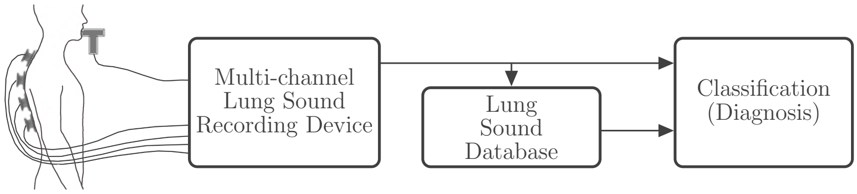

Multi-channel Lung Sound Recording Device
Lung Sound Database
Classification (Diagnosis)

*Figure 1.1: Simplified overall lung sound analysis framework.*

## 1.2 Organisation and Contributions

With the contributions summarised along the outline, this thesis is structured as follows.

- *Chapter 2:* We describe the human respiratory system and the characteristics of lung sounds from an auscultation perspective.

- *Chapter 3:* We present a robust multi-channel lung sound recording device (MLSRD). It enables 16-channel lung sound recording over the posterior chest and simultaneous airflow recording. Compared to previous approaches, we improved the usability and the robustness against air- and body-borne noise. We developed a novel lung sound transducer (LST) and an appropriate attachment method realised as a foam pad. For analogue prefiltering, preamplification, and digitization of the lung sound signal, we use a composition of low-cost standard audio recording equipment. Furthermore, we developed a suitable recording software. In simple experiments, we show the robustness of our MLSRD against ambient noise, and demonstrate the achieved signal quality. The result is an MLSRD that enables fast gathering of high-quality lung sound recordings in noisy clinical environments without impeding the daily routines.

- *Chapter 4:* We investigate the effect of airflow rate on amplitude and regional distribution of normal lung sounds. With our MLSRD, we record lung sounds over the posterior chest of lung-healthy subjects at different airflow rates. We use acoustic thoracic images to discuss the influence of the airflow rate on the regional distribution.

- *Chapter 5:* We conducted a clinical trial to record a multi-channel lung sound database for lung-healthy and pathological subjects using our MLSRD. The considered diseases are pneumothorax and idiopathic pulmonary fibrosis (IPF). This chapter contains the clinical trial design, the description of lung diseases, and the resulting database.

- *Chapter 6:* We present several neural network architectures, such as MLPs, several recurrent neural network (RNN) architectures, and convolutional neural networks (CNNs). Furthermore, we discuss some regularization methods for neural networks.

- *Chapter 7:* We present a method to accurately detect the state-sequence *first heart sound (S1) - systole - second heart sound (S2) - diastole*, i.e. the positions of S1 and S2, in single-channel heart sound recordings. We propose an event detection approach, without explicitly incorporating a priori information of the state duration. This renders it also applicable to recordings with cardiac arrhythmia and extendable to the detection of extra heart sounds (third and fourth heart sound), heart murmurs, as well as other acoustic events. We use data from the 2016 PhysioNet/CinC Challenge, containing heart sound recordings and annotations of the heart sound states. From the recordings, we extract spectral and envelope features and investigate the performance of different deep recurrent neural network (DRNN) architectures to detect the state-sequence. Our approach shows state-of-the-art performance carefully evaluated on the 2016 PhysioNet/CinC Challenge dataset.

- *Chapter 8:* We present a method for event detection in single-channel lung sound recordings. In particular, we evaluate the processing framework from Chapter 7 for the detection of crackles and breathing phase events (inspiration/expiration). For our experiments, we use the database presented in Chapter 5. The proposed method shows robustness regarding the contamination of the lung sound recordings with noise, bowel sounds, and heart sounds.

- *Chapter 9:* We present an approach for multi-channel lung sound classification, exploiting spectral, temporal and spatial information. In particular, we propose a frame-wise classification framework to process full breathing cycles of multi-channel lung sound recordings with a convolutional recurrent neural network. From the lung sound recordings, we extract spectrogram features and compare different deep neural network architectures for binary classification, i.e. lung-healthy vs. IPF. In our experiments, we use the dataset from Chapter 5. Our proposed classification framework with the convolutional recurrent neural network outperforms the other networks and achieves an F-score of $F_1 \approx 92\%$.

- *Chapter 10:* We conclude this thesis by summarising our main results and by highlighting the main characteristics and advantages of our approach. Furthermore, we discuss open research questions and future work.

- *Chapter 11:* List of publications arised during the course of my PhD studies.

# 2

# Fundamentals of Respiratory Sounds

In this chapter, we give a short introduction to respiratory sounds, organised as follows. In Section 2.1, we introduce the human respiratory system, including its function, the separation into regions, and the mechanics of breathing. An introduction to respiratory sounds from a signal characteristic prospective, including their diagnostic relevance, is then given in Section 2.2.

*This chapter is largely adopted from "Computational Fluid and Particle Dynamics in the Human Respiratory System"- Chapter 2: "The Human Respiratory System" [37], "Automatic Adventitious Respiratory Sound Analysis: A Systematic Review" [1], and from "Fundamentals of Lung Auscultation" [38].*

## 2.1 The Human Respiratory System

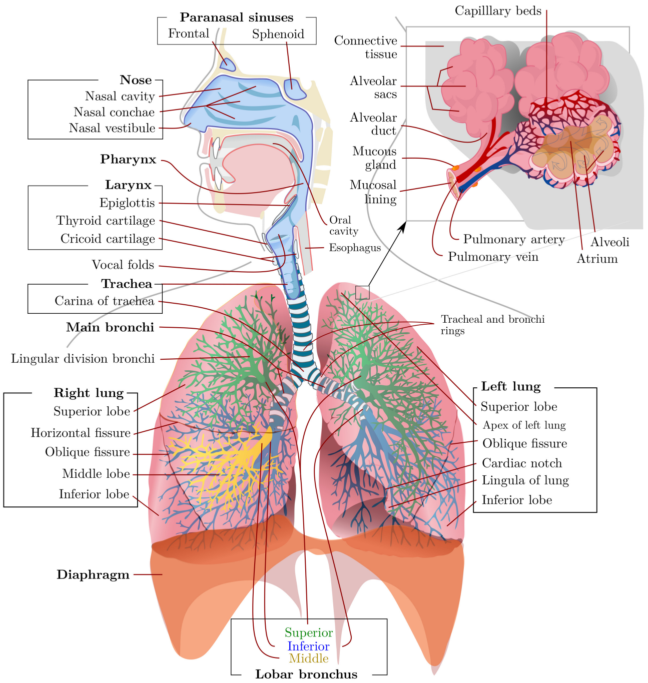


*Figure 2.1: The human respiratory system [39].*

### 2.1.1 Function

The human respiratory system consists of specific organs and structures (see Figure 2.1), which are used to supply the body with oxygen and remove carbon dioxide. It transfers oxygen from the external environment into the bloodstream, where the blood circulates it to the tissue cells. During inhalation, oxygen first enters the nose and/or mouth and passes through the larynx and the trachea to the two bronchi. Each bronchus splits into two bronchial tubes. These tubes divide into plenty of small pathways within the lung, which end in the alveoli. There, oxygen ($O_2$) is exchanged in the lung capillaries for carbon dioxide ($CO_2$). The air containing $CO_2$ is then exhaled, i.e. it returns to the bronchial pathways, the trachea and larynx, through the mouth and nose to the external environment. The respiratory system also filters, warms, and humidifies the inhaled air [37].

### 2.1.2 Separation of the Respiratory System into Regions

Two common separations of the respiratory system into regions are as follows [37]:

- *Functional Separation*:
  - *Conducting zone*: The zone from the nose to the bronchioles conducts the inhaled air to deep regions of the lungs.
  - *Respiratory zone*: Gas is exchanged in the zone from the alveolar duct to the alveoli.
- *Anatomical Separation*:
  - *Upper respiratory tract*: Contains the organs located outside of the chest cavity (thorax) area, i.e. nose, pharynx, and larynx.
  - *Lower respiratory tract*: Contains the organs located almost entirely within the chest cavity, i.e. trachea, bronchi, bronchiole, alveolar duct, alveoli.

### 2.1.3 Mechanics of Breathing

The diaphragm plays a major role in the mechanics of breathing. This muscle separates the thoracic cavity from the abdominal cavity. The act of breathing consists of the following two phases [37]:

- *Inhalation (Inspiration)*:
  During inhalation, the diaphragm contracts and descends. Muscles in the thorax pull the anterior end of each rib upwards and outwards. This increases the volume within the thorax and results in a pressure difference between the atmospheric air pressure and the inside of the thorax (i.e. intrathoracic pressure) and the lungs (i.e. intrapulmonary pressure). This leads to air flowing into the lungs. During normal breathing, the diaphragm contracts and descends about 1 cm and during forced breathing, up to 10 cm.

- *Exhalation (Expiration)*:
  Air is exhaled with the return of the lungs and chest to their equilibrium position. Within the thorax, the volume is reduced and the pressure increases. Due to this, air is released from the lungs. The elastic recoil of the lung and chest walls is sufficient to return the thorax to the equilibrium during normal breathing. For forceful breathing, additional muscles in the thorax and abdomen are needed.

## 2.2 Respiratory Sounds

Respiratory sounds are generated by the respiratory system and are usually heard with auscultation[1]. Other body sounds are heart sounds, abdomen/bowel sounds, and blood vessel sounds. Respiratory sounds are categorised as normal or abnormal, with the latter being characteristic for several lung diseases [1].

### 2.2.1 Normal Respiratory Sounds

The characteristics of normal respiratory sounds vary depending on the location where they are heard or generated. The differences are in duration, pitch, and sound quality. Normal respiratory sounds can be categorised as follows [1], with an overview presented in Table 2.1.

- *Vesicular Sounds*:
  Vesicular sounds are heard over most of the lung fields. They are audible during the whole inspiratory phase and during the early expiratory phase. Vesicular sounds from different breathing cycles are separated with a pause, but there is no separation of the sounds between inspiration and expiration within one cycle. The sounds are characterised as soft and non-musical. They have a limited frequency range between 100-1000 Hz. Due to the low-pass effect of the chest wall, there is a drop in energy after around 100-200 Hz. Sounds during inspiration are higher-pitched than those during expiration. Also, the intensity is higher during inspiration and it varies depending on the part of the chest [40, 41].

- *Bronchial Sounds*:
  Bronchial sounds are heard over the large airways near the second and third intercostal space. They are present during the inspiratory and the expiratory phase. Unlike in vesicular sounds, the bronchial sounds during expiration are present longer than during inspiration. This is due to the origin of the sounds in larger airways. Between each cycle of breathing, there is a short break. The sounds are characterised as high-pitched and tubular. Relative to vesicular sounds, they are more hollow and higher-pitched. The sound intensity during expiration is higher than during inspiration. The energy in the upper frequency range is higher than in vesicular sounds [38, 40].

[1] Auscultation is the act of listening to body sounds with a stethoscope.

- *Broncho-vesicular Sounds*:
  Broncho-vesicular sounds are heard over the posterior chest between the scapulae and centrally over the anterior chest. The inspiratory and expiratory durations are similar. The characteristics of the sounds is between vesicular and bronchial sounds. This means they are softer than bronchial sounds, but still have a tubular sound characteristic [42].

- *Tracheal Sounds*:
  Tracheal sounds are heard over the trachea. The location for auscultation is the suprasternal notch. The inspiratory and expiratory duration of tracheal sounds is similar and a distinct gap is separating them. The sound is characterised as hollow and tubular. Turbulent airflow through the pharynx and glottis is the origin of the sounds. Tracheal sounds have frequency components up to 5000 Hz, with a drop in energy after around 800 Hz. In general, they have a higher intensity than vesicular and bronchial sounds [38, 40, 43].

- *Mouth Sounds*:
  Mouth sounds are heard from the mouth. They are not audible in healthy persons. Their frequency range is from 200 to 2000 Hz, with an energy distribution similar to white noise [44].

*Table 2.1: Normal breath sounds, adopted from Table 1[*] from [1].*

| Breath Sounds | Location | Range | Pitch | Quality | Timing (I:E ratio) | Pause |
|---|---|---|---|---|---|---|
| Vesicular | Most of lung fields | 100-1000 Hz Energy drop at 200 Hz | Low | Low-pass filtered noise like Soft Rustling sound | During inspiration and early expiration (2:1 ratio) | Pause between different breath cycle |
| Broncho-Vesicular | Between scapulae on posterior chest and center part of anterior chest | Intermediate between Vesicular and Bronchial | Intermediate | Intermediate intensity | During both inspiration and expiration (1:1 ratio) | - |
| Bronchial | Large airways on chest near second and third intercostal space | Similar to Tracheal | High | Loud Hollow | During both inspiration and expiration (1:2 ratio) | Short pause between inspiration and expiration phase |
| Tracheal | Suprasternal notch on trachea | 100-5000 Hz Energy drop at 800 Hz | High | Harsh Very loud | During both inspiration and expiration (1:1 ratio) | Distinct pause between inspiration and expiration phase |
| Mouth | Mouth | 200-2000 Hz Low | High | White-noise like Silent when normal | - | - |

[*] `https://doi.org/10.1371/journal.pone.0177926.t001`

### 2.2.2 Abnormal Respiratory Sounds

Abnormalities in respiratory sounds during breathing can take the form of a decrease in sound intensity, the presence of normal sounds in abnormal areas, or adventitious sounds [1].

### Decrease in Sound Intensity

The most common abnormality is the decrease in sound intensity. This can be due to the following two reasons [38]:

- *Decreased sound energy at the site of generation*:
  A decrease in sound energy can result from a drop in inspiratory airflow, caused by:
  - Poor cooperation of the patient (e.g. unwillingness to take a deep breath).
  - Depression of the central nervous system (e.g. drug overdose).
  - Blockage of the airways (e.g. by a foreign body or tumor).
  - Narrowing due to obstructive airway diseases (e.g. asthma and chronic obstructive pulmonary disease (COPD)).

  The decrease in intensity can be
  - permanent (e.g. emphysema) or
  - reversible (e.g. asthma attack).

- *Impaired transmission*:
  - *Intrapulmonary factors*:
    - Disruption of the mechanical properties of the lung parenchyma (e.g. a combination of hyperdistention and parenchymal destruction in emphysema).
    - A medium with a different acoustic impedance than the normal parenchyma is found between the source of sound generation and the stethoscope (e.g. collections of gas or liquid in the pleural space - pneumothorax, hemothorax, and intrapulmonary masses).
  - *Extrapulmonary factors*:
    - Obesity.
    - Chest deformities.
    - Abdominal distention due to ascites.

### Presence of Normal Lung Sounds in Abnormal Areas

An example for the presence of normal lung sounds in abnormal areas is *bronchial breathing.* Bronchial sounds are abnormal if they are heard over the peripheral areas of the lung. This can be caused by the development of lung consolidation in pneumonia. The blocking of the embedded airways due to inflammation or viscous secretions causes decreased lung sounds. However, in patent airways the sound transmission is improved. This increases the expiratory component, resulting in *bronchial breathing* [38].

### Adventitious Sounds

In some lung conditions, adventitious sounds are superimposed on normal breath sounds. The characteristics of adventitious sounds are related to the underlying disease. Depending on the duration, it is distinguished between continuous adventitious sounds (CAS) and discontinuous adventitious sounds (DAS) as follows [1]. A short summary is also presented in Table 2.2.

- *Continuous Adventitious Sounds (CAS)*:
  CAS are adventitious sounds with a duration longer than 250 ms [45]. Due to the associated condition and cause, they can be classified as follows:
  - *Wheeze and Rhonchi*:
    Wheeze and rhonchi have a musical sound with a frequency range between 100-1000 Hz. They are sinusoid-like signals with up to three harmonic frequencies. Although categorised as CAS, they can be shorter than 250 ms, with minimum duration of 80 to 100 ms. Both sounds are audible during inspiration and/or expiration [38, 46, 47]. The differences are as follows:
    - *Wheeze*:
      Wheezes are continuous sounds with a pitch higher than 400 Hz. They are caused by airflow limitation due to airway narrowing. Related diseases are asthma, COPD, or even a tumour [38, 45, 48].
    - *Rhonchi*:
      Rhonchi are continuous sounds with a maximal pitch of 200 Hz. The sound is caused by thickening of mucus in the larger airways. Related diseases are COPD and bronchitis [38, 45, 49].
  - *Stridor*:
    Stridors sound sibilant and musical. They are similar as wheezes, but normally harsher and louder. Their duration is longer than 250 ms and they have a pitch higher than 500 Hz. They can be heard during inspiration and expiration. Because stridors are caused by turbulent airflow in the larynx or bronchial tree, they are more clearly heard on the trachea. They are related to upper airway obstruction, like epiglottitis, croup, and laryngeal oedema. Also a foreign body, like a tumour in the upper airways, can cause stridors [38, 50].
  - *Gasp*:
    Gasps are characterised as having a whoop-like sound. They are audible during inspiration after a bout of coughing. The origin is fast movement of air in the respiratory tract. They are related to pertussis, also known as whooping cough [51].
  - *Squawk*:
    Squawks are sometimes called *short wheezes*, because they are similar as lower-pitched wheezes, but shorter in duration. Their character is a mix of musical and non-musical. The pitch is around 200-300 Hz and they appear during inspiration. The origin of the sounds is due to oscillation at the peripheral airways. They are associated with hypersensitivity pneumonia and common pneumonia [38, 40, 52, 53].

- *Discontinuous Adventitious Sounds (DAS)*:
  DAS have a duration shorter than 25 ms. Based on the source of sound generation, they can be classified as follows [1]:
  - *Fine Crackles*:
    Fine crackles are characterised as explosive and non-musical. Their duration is around 5 ms and their pitch around 650 Hz. They appear only late during inspiration. Fine crackles are generated with the explosive opening of the small airways. They appear in pneumonia, congestive heart failure, and IPF [40, 54, 55].
  - *Coarse Crackle*:
    Compared to fine crackles, coarse crackles are low-pitched, with a frequency around 350 Hz and a duration of around 15 ms. They are audible mostly early during inspiration, but also during expiration. The sounds are caused by air bubbles in large bronchi. Related diseases are chronic bronchitis, bronchiectasis, and COPD [54].
  - *Pleural Rub*:
    Pleural rubs are characterised as non-musical and rhythmic. Their duration is around 15 ms and their pitch usually below 350 Hz. They are audible during inspiration and expiration. Pleural rub is generated with the rubbing of the pleural membranes during breathing. Related conditions are the inflammation of the pleural membrane or a pleural tumor [38, 40].

### 2.2.3 Usage of the Terms Breath Sounds, Adventitious Sounds, Lung Sounds and Respiratory Sounds with Respect to the Anatomical Location

Figure 2.2 illustrates the connection of the terms breath sounds, adventitious sounds, lung sounds and respiratory sounds. The distinction is based on the anatomical location of sound recording and the composition of the sounds, i.e. with or without adventitious sounds [56].

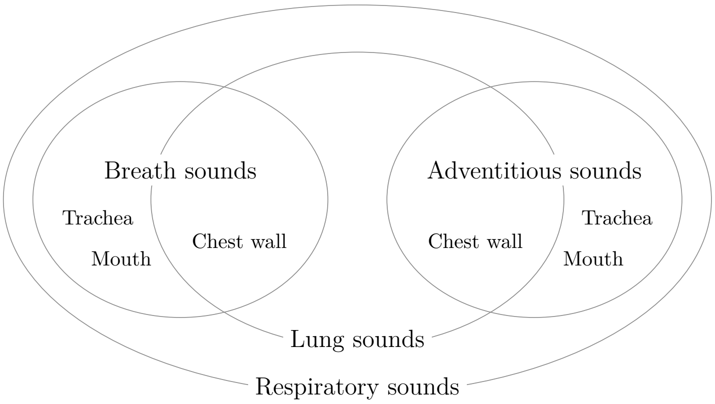


*Figure 2.2: Connection of the terms breath sounds, adventitious sounds, lung sounds and respiratory sounds [56].*

*Table 2.2: Adventitious sounds and their characteristics, adopted from Table 2* from [1].*

| Types | Continuity | Duration | Timing | Pitch | Quality | Cause | Disease Associated |
|---|---|---|---|---|---|---|---|
| Wheeze | Continuous | $> 80$ ms | Inspiratory, Mostly Expiratory, Biphasic | High ($> 400$ Hz) | Sibilant, Musical | Airway narrowing, airflow limitation | Asthma, COPD, Foreign body |
| Rhonchi | Continuous | $> 80$ ms | Inspiratory, Mostly Expiratory, Biphasic | Low ($< 200$ Hz) | Sibilant, Musical | Secretion in bronchial, muchosal thickening | Bronchitis, COPD |
| Stridor | Continuous | $> 250$ ms | Mostly Inspiratory, Expiratory, Both | High ($> 500$ Hz) | Sibilant, Musical | Turbulent airflow in larynx or lower bronchial tree (Upper airway obstruction) | Epiglottitis, foreign body, croup, laryngeal oedema |
| Fine Crackle | Discontinuous | ± 5 ms | Inspiratory (late) | High (650 Hz) | Non-musical, Explosive | Explosive opening of small airways | Pneumonia, Congestive heart failure, Lung fibrosis |
| Coarse Crackle | Discontinuous | ± 15 ms | Mostly Inspiratory (early), Expiratory, Both | Low (350 Hz) | Non-musical, Explosive | Air bubble in large bronchi or bronchiectatic segments | Chronic bronchitis, bronchiectasis, COPD |
| Pleural Rub | Discontinuous | $> 15$ ms | Biphasic | Low ($< 350$ Hz) | Non-Musical, Rhythmic | Pleural membrane rubbing against each other | Inflammation of lung membrane, lung tumour |
| Squawk | Continuous | ± 200 ms | Inspiratory | Low (200—300 Hz) | Short Musical and non-musical | Oscillation of peripheral airways | Hypersensitivity pneumonia, pneumonia |
| Gasp | Continuous | $> 250$ ms | Inspiratory | High | Whoop | Gasping for breath | Whooping cough |

*https://doi.org/10.1371/journal.pone.0177926.t002

# 3 Multi-channel Lung Sound Recording Device

In this chapter, we present a robust multi-channel lung sound recording device (MLSRD). It enables 16-channel lung sound recording over the posterior chest and simultaneous airflow recording. We developed a novel LST and an appropriate attachment method realised as a foam pad, our so called auscultation pad. Compared to previous approaches, the usability and robustness against air- and body-borne noise is greatly improved.

This chapter is organised as follows. After a short introduction in Section 3.1, we present our LST design and the auscultation pad in Section 3.2. Section 3.3 gives an overview on the remaining components of the MLSRD. The achieved signal quality and the robustness against ambient noise is discussed in Section 3.4. Section 3.5 describes the measurement procedure and Section 3.6 provides information about the medical safety. We conclude this chapter in Section 3.7.

*The main parts of this chapter were published in the conference proceedings of the 9th International Conference on Biomedical Electronics and Devices 2016, as "A Robust Multichannel Lung Sound Recording Device" [57]. As minor modifications, we changed some wordings, added some sentences and figures, and inserted sections for signal level calibration and medical safety.*

## 3.1 Introduction

Sensors applied to lung sound recording are contact sensors or air-coupled microphones. The latter are either microphones with custom designed couplers or microphones inserted into stethoscope rubber tubes [9, 10]. The most common recording technique employs air-coupled microphones attached with self-adhesive tape. However, this approach lacks of sensitivity against body sounds and ambient noise [58–60]. Moreover, for multi-channel usage the attachment of several LSTs with self-adhesive tape results in a poor usability and increases the sensitivity to measurement errors.

To circumvent the afore-mentioned drawbacks, we introduce a robust MLSRD. It supports to record a high quality lung sound database for multi-channel lung sound classification. To obtain clean lung sound recordings, we focused on the recording stage and reduced post-processing. Furthermore, it was important that the MLSRD is suitable to record lung sounds for a large number of diseases. This is reflected in the distribution and position of the LSTs. Besides distinct adventitious lung sounds [61], it should reliably allow the estimate of changes in amplitude of lung sounds, which is necessary for the detection of, e.g. pneumothorax [62].

Based on the approach with air-coupled microphones [9], we developed a novel lung sound transducer to ensure a high signal quality. For the attachment of the LSTs, we designed a foam pad similar to the Stethographics STG 16 [11]. It records lung sounds over the posterior chest in supine position. We implemented the analog prefiltering, preamplification, and digitization of the lung sound signal with a composition of standard audio recording equipment. The entire MLSRD consists of the foam pad (auscultation pad) and a pneumotachograph, both working with an appropriate recording software on a personal computer.

## 3.2 Multi-channel Recording Front-end

The core of our MLSRD is the auscultation pad. It is a foam pad with 16 LSTs distributed on its surface. We adapted our LST design for this attachment method. In the following subsections, we separately describe the LST design and the foam pad.

### 3.2.1 Lung Sound Transducer

We developed a novel LST according to the approach with air-coupled electret-condenser microphones [9]. It is shown in Figure 3.1. We use a Littmann Classic II chest piece as coupler. We inserted an electret-condenser microphone capsule (ECMC) in such a way that the stethoscope chest piece is acting as a conical coupler between the microphone capsule and the human skin. The depth of the conical coupler is $d = 2.2$ mm, and the width is $w = 33$ mm. Its shape corresponds with the recommendations in [63, 64]. We used the diaphragm of the chest piece to cover its opening. It prevents the filling of the coupler cavity with skin, and, thus, it ensures its acoustic effect. This is important for varying contact pressure, and, therefore, it is relevant for our attachment method discussed in Section 3.2.2. The diaphragm further enables a convenient disinfection of the LST, and it protects the ECMC from scratching body hair and dirt.

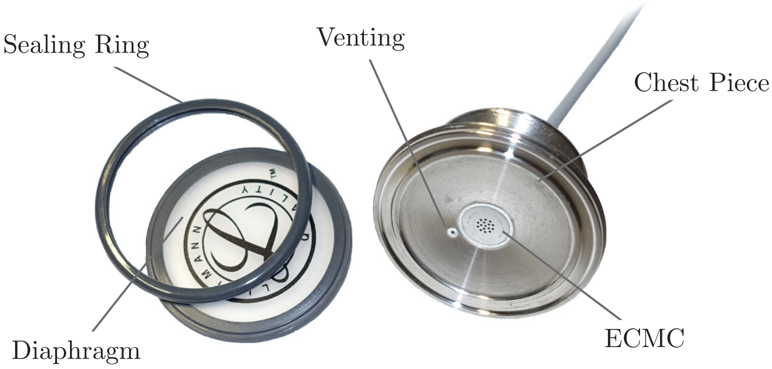


*Figure 3.1: Lung sound transducer consisting of an ECMC inserted in a Littmann Classic II chest piece.*

To allow static pressure equalization between the coupler chamber and the surrounding air, we inserted a small vent into the chest piece. We used a thin-wall 23-g needle with a length of $l = 11.5$ mm and with an inner diameter of $d = 0.35$ mm, according to the recommendations in [64]. As ECMC, we used the Primo EM172, which features a very high signal-to-noise ratio of $SNR = 80$ dB and a sensitivity of $-27$ dB (re $1V/Pa$). These specifications are distinctly better than those of widely-used microphones, like the Sony ECM-44BPT [12] or the Sony ECM-77B [6], which feature an $SNR \leq 64$ dB.

### 3.2.2 Auscultation Pad

The attachment of the LST is crucial, because of its high sensitivity against air- and tissue-borne sounds [58, 59]. Therefore, we developed a foam pad, the auscultation pad shown in Figure 3.2. It consists of several foam layers and a cover of artificial leather. The topmost layers holds the LSTs. There is a small cavity beneath each LST to avoid increasing contact pressure due to the underlying foam. Furthermore, the cavity prevents the foam from touching or even clogging the venting of the LST. By using different kinds of foam, we designed a shape that adapts to almost every physique. This construction provides a symmetric contact pressure with respect to the spine. We arranged the LSTs on the surface of the auscultation pad with a fixed pattern, which almost matches the one proposed in [12]. The pad enables a fast attachment of the LSTs on the posterior chest by simply placing the auscultation pad under the back of the patient in supine position. To further stabilize the patient, we extended the auscultation pad with two additional pads, one for the head and another one for the buttocks, as shown in Figure 3.3. We are able to

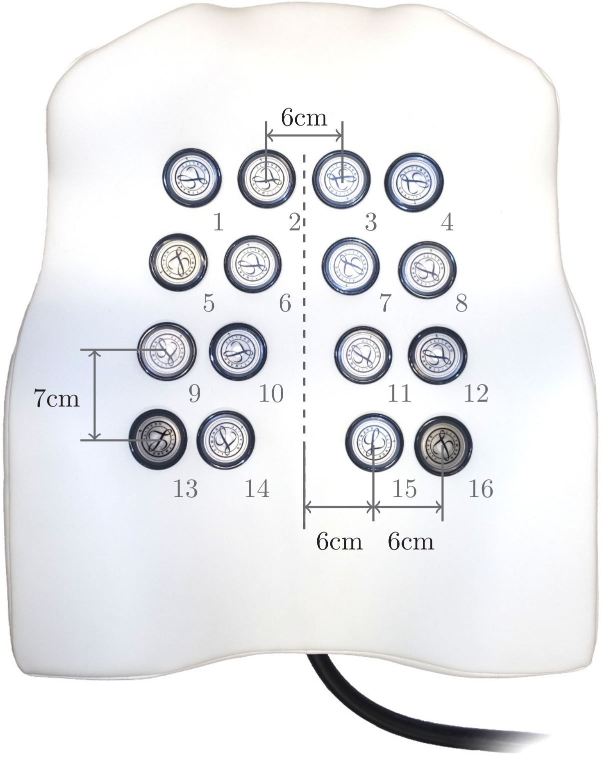


*Figure 3.2: 16-channel lung sound recording front-end.*

achieve a high robustness against air- and body-borne noise with an overall high lung sound signal quality. Further details are presented in Section 3.4. The attenuation of ambient noise is due to the surrounding foam. We achieve the robustness against body-borne noise due to the reliable attachment and almost no movement of the back during breathing in the supine position. The surrounding foam further protects the LST cable from body-borne noise. Another advantage is the balanced audio connection from the auscultation pad to the microphone preamplifiers, which reduces the susceptibility to external noise.

## 3.3 Multi-channel Lung Sound Recording Device

We use the auscultation pad as part of a mobile recording setup (Figure 3.3). The setup consists of an equipment cart, which includes the recording hardware, two screens, a pneumotachograph, and the pads. The following subsections contain some details about the remaining components.

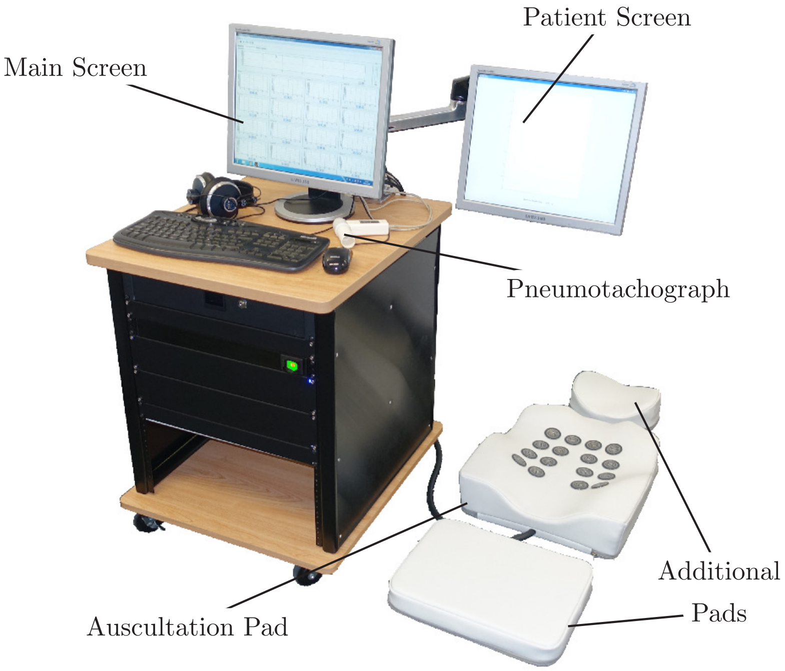


*Figure 3.3: Mobile lung sound recording device containing the auscultation pad and the additional components.*

### 3.3.1 Recording Hardware

We implemented the analogue prefiltering, preamplification, and digitization of the lung sound signal with low-cost standard audio recording equipment. The composition of the appropriate hardware fulfills the requirements of the computerised respiratory sound analysis (CORSA) guidelines [65].

We use two SM Pro Audio EP84 8-channel microphone preamplifiers with the integrated ADAT interface SM Pro Audio PR8IIA. In addition to high-pass filtering (cut-off frequency $f_c = 80$ Hz, with a slope of 18 dB/oct), preamplification, and analog-to-digital conversion of the LST signal, the SM Pro Audio PR8IIA also provides the supply voltage (phantom power) for the ECMCs. For the suitable operating voltage of the ECMCs, we further use AKG MPA VL phantom power adapters. They convert the phantom power of 48V to the required 3V∼10V. The AKG MPA VL phantom power adapter features a high-pass characteristic with a cut-off frequency of $f_c = 80$ Hz and a slope of 6 dB/oct. In a series-connection with the microphone preamplifier, an overall high-pass characteristic with a slope of 24 dB/oct is achieved. The high-pass filters of the microphone preamplifiers and the phantom power adapters are implemented as Bessel filters. Therefore, they feature an approximately linear phase response. The two SM Pro Audio EP84 are connected with an RME Fireface 800 audio interface. This represents a firewire multi-channel audio recording device for a computer.

In short, for analogue prefiltering, we apply a Bessel high-pass with a cut-off frequency of $f_c = 80$ Hz and a slope of 24 dB/oct. For analog-to-digital conversion, we us a sampling frequency of $f_s = 48$ kHz and a resolution of 24 bit. Before storing, we resampled the audio files to $f_s = 16$ kHz, where an anti-aliasing low-pass with cut-off frequency of $f_c = 7.6$ kHz is applied.

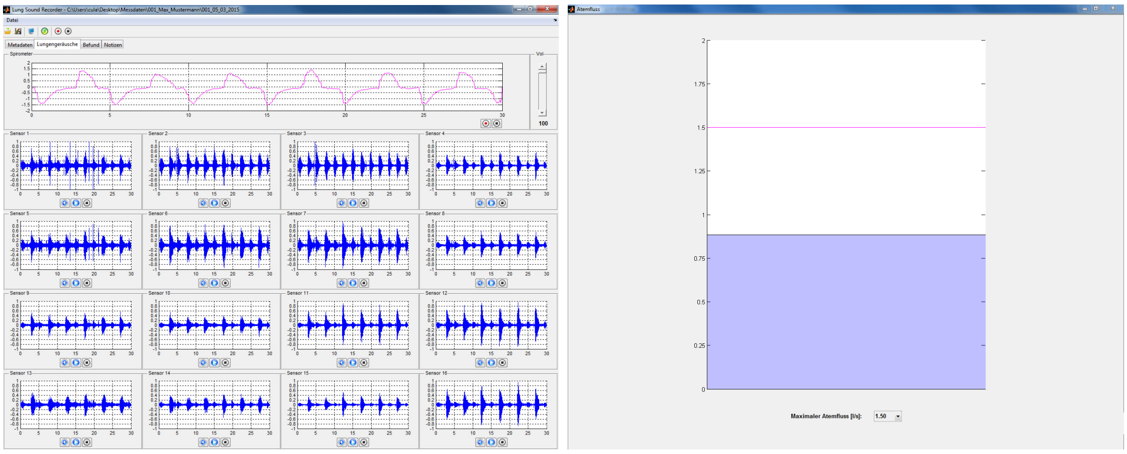
The high-pass filter is applied to reduce low-frequency distortion to the signal. Possible distortions are heart sounds, muscle noise, external low-frequency noise, and noise due to changes in contact pressure (caused by body or sensor motion) [65].

### 3.3.2 Airflow Recording - Pneumotachograph

The simultaneous recording of the velocity of respired air and the lung sounds makes the distinction between inspiratory and expiratory phases possible. Furthermore, we almost reach a uniform lung sound signal intensity profile by specifying the respiratory behaviour of the patient, resulting in a high quality database. We use a Schiller SP 260 pneumotachograph connected via the USB port. The sampling frequency for the airflow signal is $f_s = 400$ Hz.

### 3.3.3 Recording Software

We developed a MATLAB graphical user interface (GUI) for the simultaneous recording of the airflow signal and the lung sounds. We use Playrec [66] for the multi-channel recording of the lung sound signals with MATLAB. For the flow signal, we read the serial port of the pneumotachograph. Figure 3.4a shows the main screen of the software, illustrating the simultaneously recorded airflow and the lung sounds. Figure 3.4b shows a patient screen to display the measured absolute value of the air flow in real-time.



Lung Sound Recorder - C:\Users\cula\Desktop\Messdaten\001_Max_Mustermann\001_05_03_2015

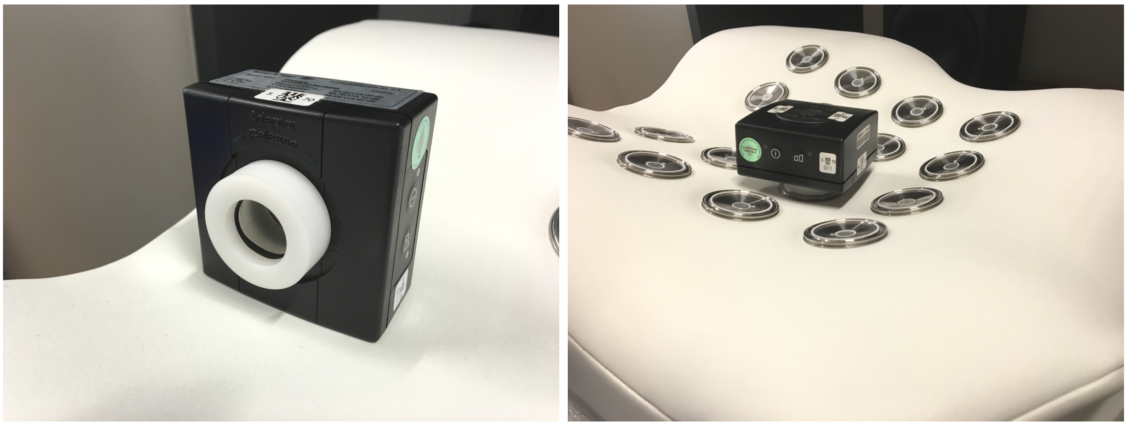
Maximaler Atemfluss [l/s]: 1.50

(a) (b)

*Figure 3.4: MATLAB GUI: a) shows the main screen with recordings of the airflow signal and the lung sounds and b) the patient screen for real-time feedback for the airflow.*

### 3.3.4 Signal Level Calibration

For the calibration of the recording device, we used a Brüel & Kjær sound calibrator Type 4231, as depicted in Figure 3.5. For this purpose, we first detached the diaphragms of the LSTs. The calibrator was then attached with a suitable adapter. We adjusted the microphone preamplifiers of the LSTs to reach the same signal level for the sound calibrator signal, which is a sinusoidal waveform with a frequency of $f = 1$ kHz and a sound pressure level of 94 dB.



(a) (b)

*Figure 3.5: Calibration with a Brüel & Kjær sound calibrator Type 4231: a) shows the calibrator including the adapter and b) its attachment to a LST.*

## 3.4 Robustness and Signal Quality

In this section, we show the achieved signal quality by means of the signal-to-noise ratio (SNR) and the frequency range of the recording setup. We further demonstrate the robustness against ambient noise with a simple experiment.

### 3.4.1 Robustness Against Ambient Noise

The CORSA guidelines [67] recommend to have environmental condition with a background noise level of preferably $< 45$ dB$_A$ during lung sound recordings. These requirements are not always given in clinical settings.

We compared the performance of our auscultation pad in a noisy setting with the attachment-method of the LST with self-adhesive tape [9]. The experiment took place in a small room. We considered two measurement scenarios. In the first scenario, we centered the auscultation pad on the floor with a (male) test person lying on it. In the second scenario, we placed a chair in the center of the room, with the test person in a sitting position. With self-adhesive tape, we attach an LST on the persons posterior chest, at the same position where it was attached with the auscultation pad. The rest of the signal acquisition chain remained the same for both scenarios, as introduced in Section 3.3. As noise sources, we used five loudspeakers, which played back white Gaussian noise. The loudspeakers feature a flat frequency response from $f = 80$ Hz to $f = 20$ kHz. We measured the A-weighted equivalent sound level for both scenarios at the position of the sensor over 30 seconds with $L_{Aeq} = 68$ dB. During the recording of the LST signal, we instructed the test person to hold his breath for 15 seconds and played back the white Gaussian noise.

Figure 3.6 shows the power spectral density (PSD) for the LST signal in both scenarios in the relevant frequency range. We see a distinct attenuation compared to the attachment with self-adhesive tape, starting above a frequency of $f = 500$ Hz. At a frequency of $f \approx 2$ kHz the difference is up to 50 dB.

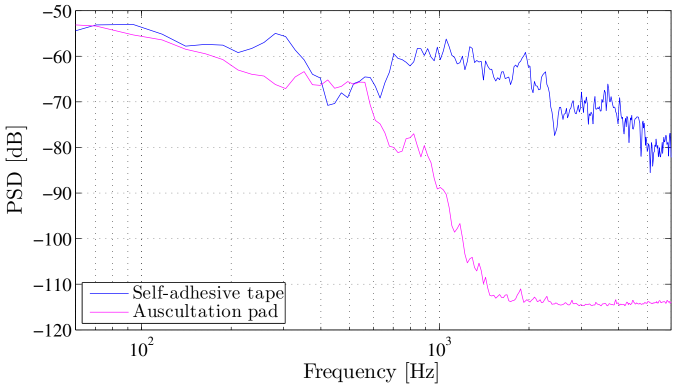


*Figure 3.6: Attenuation of background noise of the auscultation pad compared to the attachment method of lung sound transducers with self adhesive tape.*

### 3.4.2 Signal Quality

In Figure 3.7, we show the SNR of our recording setup by illustrating the spectral characteristics of a vesicular lung sound of a healthy adult person. The blue line shows the spectral characteristics during the inspiratory phase. The red line shows the background noise recorded at breath hold; the frequency components in the low-frequency range are mainly caused by body movement and heart sounds. We achieve a signal-to-noise ratio of up to $SNR \approx 40$ dB in the relevant frequency range. Due to the high SNR of the microphone, we observe in this vesicular lung sound recording frequency components up to $f \approx 2.5$ kHz.

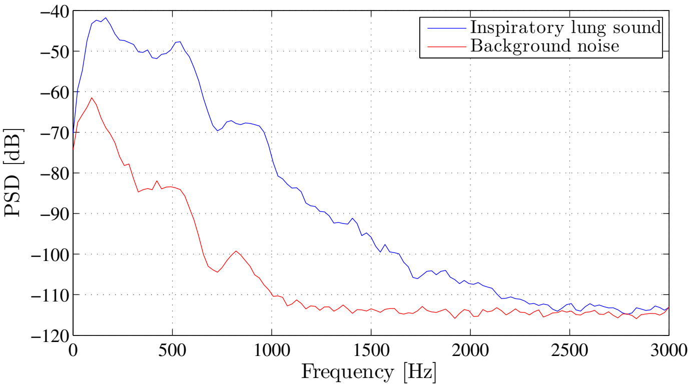


*Figure 3.7: Spectral characteristics of normal breath sounds and background noise at breath hold, recorded over the posterior chest of a healthy adult.*

## 3.5 Measurement Procedure

The lung sounds are recorded in supine position on an examination table. The auscultation pad is placed under the back of the subject. For the orientation of the subject on the pad, we use a defined distance $d$ between the 7th cervical (C7) vertebra and the topmost row of sensors, as illustrated in Figure 3.8. The size of the lung varies depending on the subject's physique. The different sized lungs result in a scaling of the organ along the sensor grid towards C7. For small and thin subjects, the lower and/or outer sensors can become irrelevant, i.e. not all sensors are needed to cover the lung. During a measurement, the subject is instructed to lie quietly on the auscultation pad, to hold the pneumotachograph with both hands, and to wear a nose clip. The breathing behaviour is specified with a maximum airflow value. This value should be reached during inspiration with a natural breathing behaviour during expiration. A real-time feedback for the airflow is provided on the patient screen.

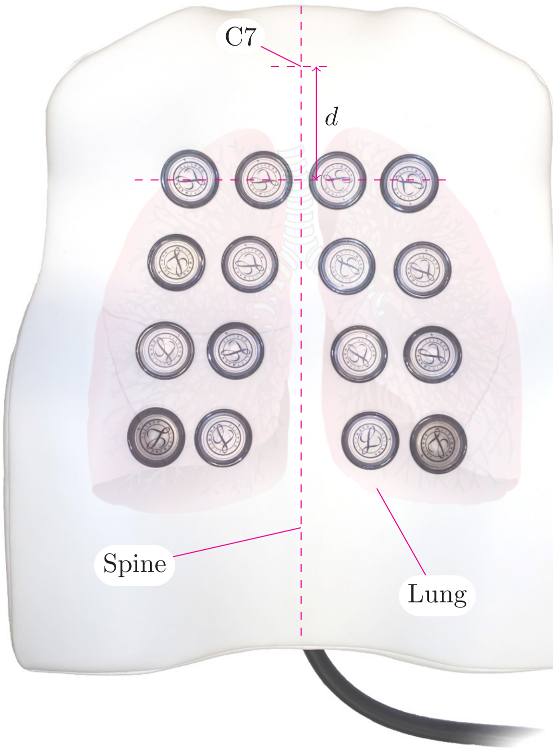


*Figure 3.8: 16-channel lung sound recording front-end with illustrated position of the sensors over the lung.*

## 3.6 Medical Safety

The basic safety requirements of the MLSRD according to European Standard (EN) 60601-1/2006+A1/2013 has been verified by the European testing and certification body for medical devices (No. 0636) at Graz University of Technology.

## 3.7 Conclusion

We developed a robust MLSRD, which reliably records high quality lung sounds. With simple measurements, we successfully underline the robustness of our auscultation pad with respect to ambient noise. Compared to the attachment of the LST with self-adhesive tape, we achieve an attenuation of ambient noise of up to 50 dB in the relevant frequency range. Due to the high signal-to-noise ratio of our LST's microphone of $SNR = 80$ dB, we obtain a bandwidth of up to $f = 2500$ Hz for vesicular lung sounds. Our MLSRD allows the recording of lung sounds in real clinical settings without the need for postprocessing, e.g. signal denoising for background speech removal.

# 4 Impact of Airflow Rate on Normal Lung Sounds

In this chapter, we investigate the effect of airflow rate on amplitude and regional distribution of normal lung sounds. We analysed and evaluated the properties of multi-channel lung sound recordings and the device itself, in order to provide information for the clinical trial design. This includes information about the interpretability of decrease in lung sound intensity and the relevance of airflow rate during data collection. Additionally, we gained experience on how to calibrate the recording device, i.e. how to adjust the pre-amplifiers to avoid clipping of the audio signals.

This chapter is organised as follows. We give a short introduction in Section 4.1. Section 4.2 describes the data acquisition, the subjects and the recording material, the signal energy calculation, and the generation of acoustic thoracic images. Section 4.3 presents our observations for the regional distribution and the lung sound amplitude for different airflow rates. In Section 4.4, we discuss the results and conclude this chapter.

*This chapter was published in the conference proceedings of the 10th Annual International conference on Bio-inspired Systems and Signal Processing 2017 under the title of "Impact of Airflow Rate on Amplitude and Regional Distribution of Normal Lung Sounds" [68]. As minor modifications, we changed some wordings, and added some sentences and a new figure.*

## 4.1 Introduction

In auscultation, besides distinct findings like adventitious lung sounds, the lung sound intensity is also used as a diagnostic marker. For example, physicians examine the differences in intensity between left- and right-sided lung sounds at pneumothorax condition. Therefore, basic knowledge about the regional distribution of normal lung sound intensity, as well as its dependence on airflow rate is essential. Moreover, a good understanding of this dependence could render the pneumotachograph dispensable for lung sound research, because airflow could be estimated directly from lung sounds [69].

Several research groups previously investigated the effect of airflow rate on the amplitude and the regional distribution of lung sound. Differing relationships were observed in [70], [71], [72], and [73]. Recently, the authors in [74] showed the effect of airflow rate on vibration response imaging measurement in healthy lungs during expiration, but also discussed the relationship between lung sound energy and airflow rate. The authors in [75] used a 5x5 microphone array and generated respiratory acoustic thoracic images (RATHIs) to discuss the regional distribution of lung sounds, by comparing its performance with clinical physicians. In [76], the authors further show RATHIs at different airflow rates.

Within this chapter, we independently investigate the impact of airflow rate on amplitude and regional distribution of normal lung sounds. For that, we recorded lung sounds on the posterior chest of four lung-healthy male subjects with our 16-channel lung sound recording device (see Chapter 3) at airflow rates of 0.3, 0.7, 1.0, 1.3 and 1.7 l/s during inspiration. In contrast to other research groups, we recorded lung sounds in supine position. Another difference is the usage of uncontaminated lung sound recordings, i.e. free of heart and other interfering sounds. By means of acoustic thoracic images [77], we discuss the regional distribution of lung sounds depending on airflow rate. To generate the surface acoustic thoracic images from the multiple lung sound signals, we use 2D-interpolation. For each subject, we illustrate the acoustic thoracic images at the five airflow rates independently.

## 4.2 Materials and Methods

### 4.2.1 Subjects and Material

At airflow rates of 0.3, 0.7, 1.0, 1.3 and 1.7 l/s, we recorded lung sounds over the posterior chest of four lung-healthy subjects. The subjects held the pneumotachograph with both hands and wore a nose clip. The subjects were instructed to breathe steadily during inspiration at the given airflow rate and with natural breathing during expiration. The subjects were placed on the pad with a defined distance $d \approx 7$ cm between the 7th cervical vertebra (C7) and the center line of the topmost row of sensors. Figure 4.1 shows examples of phonopneumograms (overlapping illustration of lung sounds and airflow signals) for one subject, recorded with sensor 12 [69]. The recording material of one subject consists of 16-channel lung sound recordings at five different airflow rates, with 4-8 breathing cycles within 30 seconds, respectively. The subjects were four male volunteers, with no diagnosed lung diseases and with the following metadata: age (27, 27, 26, 27 years), weight (78, 75, 75, 75 kg) and height (1.8, 1.78, 1.89, 1.72 m).

Our multi-channel recording front-end is robust against ambient noise. However, in lung sound recordings, interfering signals are caused by the heart, bowels, and body movement. These can distort the signal energy values from lung sound signals. To ensure uncontaminated lung sound recordings, we manually labeled the sections containing heart, bowel and other interfering sounds.

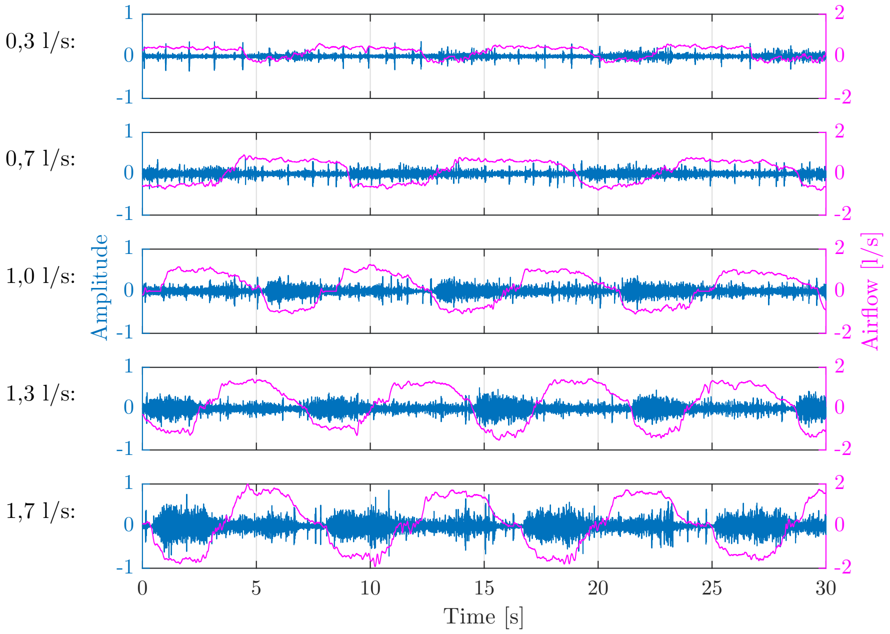


*Figure 4.1: Phonopneumograms of sensor 12 from one subject for different maximum inspiratory airflow values (0.3, 0.7, 1.0, 1.3 and 1.7 l/s) [69].*
*Legend: — Lung Sound Recording; — Airflow of Pneumotachograph*

### 4.2.2 Signal Energy Calculation

We applied a bandpass filter, with a lower cut-off frequency $f_L = 150$ Hz and an upper cut-off frequency $f_H = 250$ Hz, to the 16 lung sound signals. To calculate the energy, we used a sliding window with a length of 50 ms and an overlap of 75 %.

### 4.2.3 Acoustic Thoracic Images

To illustrate the regional distribution of the lung sound energy, we use acoustic thoracic images. We generate the images for the left and the right hemithorax independently. In particular, we use the energy signal of the left-sided (Sensors 3, 4, 7, 8, 11, 12, 15 and 16) and right-sided sensors (Sensors 1, 2, 5, 6, 9, 10, 13 and 14), respectively. To generate an acoustic thoracic image at a certain airflow rate, we used the appropriate segments of the recording. We average the energy values of all the uncontaminated segments, i.e. labeled as free of interfering sounds (cf. Section 4.2.1), where the subjects reached the proper airflow rate. For the interpolation between the energy values, which we obtained from the eight sensor signals, we use the biharmonic spline interpolation. This results in grayscale acoustic thoracic images (cf. Figure 4.4). The white color indicates the minimum value and the black color the maximum value.

## 4.3 Results

### 4.3.1 Amplitude

Figure 4.2 shows the square root of the sound energy as a function of airflow rate for all of the four subjects independently. We performed linear regression for the values from each subject independently. The coefficients of determination are $R^2 = [0.98, 0.96, 0.99, 0.99]$ for Subject 1, Subject 2, Subject 3, and Subject 4, respectively.

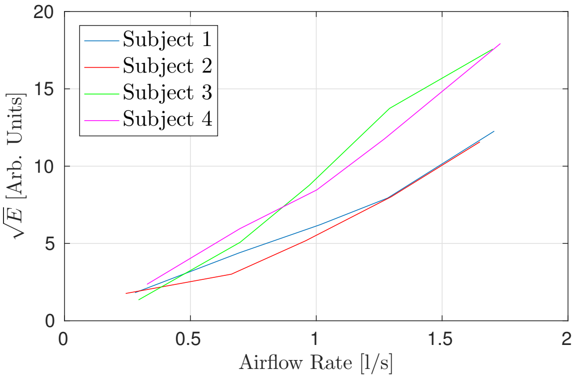


*Figure 4.2: Square root of the sound energy $\sqrt{E}$ as a function of airflow rate for all four subjects.*

Figure 4.3 shows the spectral characteristics (i.e. PSD) of the lung sounds at different airflow rates, generated from the lung sound recording of Sensor 6 (see Figure 3.2) from Subject 1. With increasing airflow rate the signal energy increases, resulting in a better SNR. The limiting factor in the higher frequency range is the noise floor of the microphone. For an airflow rate of 0.3 l/s, we observe frequency components up to $f \approx 1.2$ kHz. For an airflow rate of 1.7 l/s, we observe frequency components up to $f \approx 2.5$ kHz.

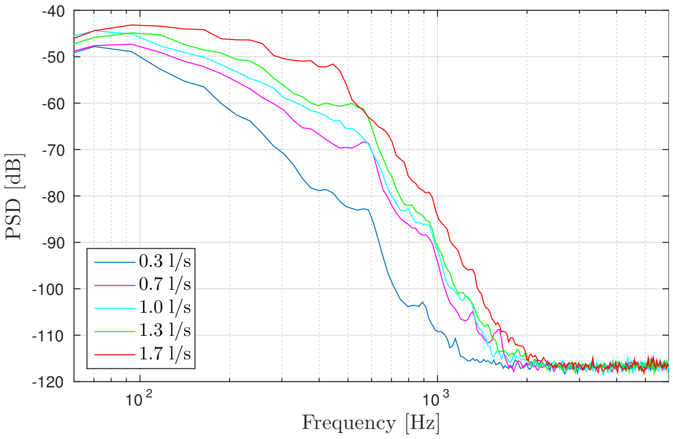


*Figure 4.3: Spectral characteristics of the lung sounds at different airflow rates, generated from the lung sound recording of Sensor 6 (see Figure 3.2) from Subject 1.*

### 4.3.2 Regional Distribution

Figure 4.4 shows the acoustic thoracic images of the four subjects, evaluated at five different airflow rates. In each acoustic thoracic image, the white and black color indicate the lowest and highest energy value of the respective recording. For an airflow rate of 0.3 l/s, we observe that most of the energy is in the middle right area. Already for an airflow rate of 0.7 l/s, the lung sound energy is higher towards the base of the lungs. Above an airflow rate of 0.7 l/s, the regional distribution remains highly similar.

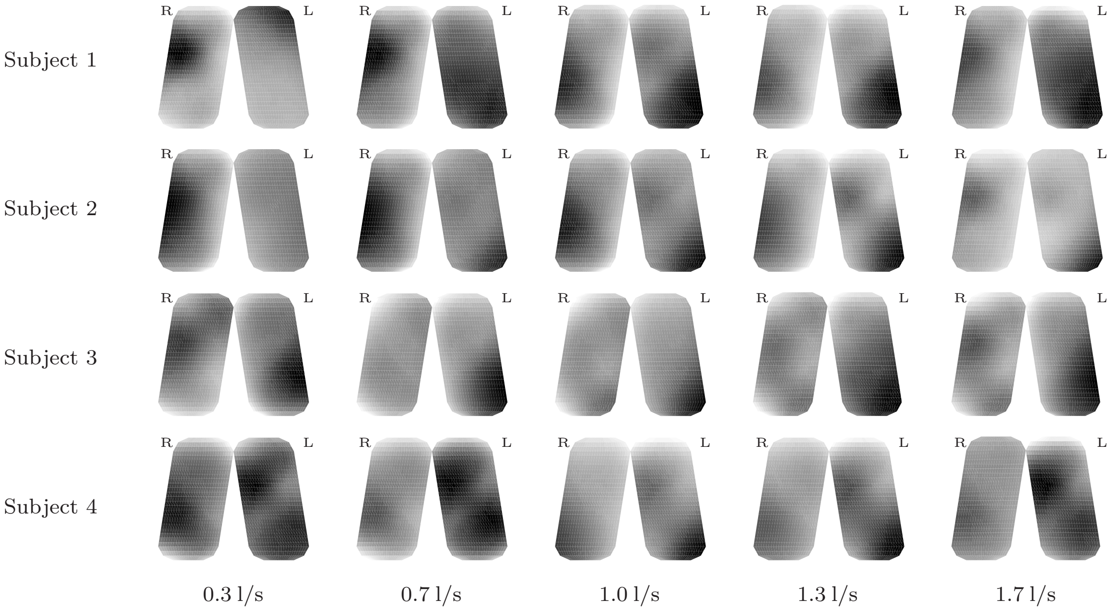


*Figure 4.4: Acoustic thoracic images from four lung-healthy subjects, evaluated at five different airflow rates. The orientation is indicated by the capital letters R (right hemithorax) and L (left hemithorax).*

Table 4.1 shows the energy distribution over either left and right, or upper and lower hemithorax. For this, we summed up the energy from the eight sensors over the respective hemithorax at the different airflow values, respectively. We report the mean and standard deviation from the four subjects. The signal energy over the left hemithorax is distinctly higher than over the right one. This is further reflected in the acoustic thoracic images, especially above an airflow rate of 0.3 l/s. We also observe that with increasing airflow rate the energy over the left hemithorax increases. Regarding the ratio of the upper to lower hemithorax, for an airflow value of 0.3 l/s the energy in the upper one is higher. With increasing airflow, the value for the lower hemithorax increases, but for 1.7 l/s it decreases again.

*Table 4.1: Energy distribution over either left and right, or upper and lower hemithorax.*

| | 0.3 l/s | 0.7 l/s | 1.0 l/s | 1.3 l/s | 1.7 l/s |
|---|---|---|---|---|---|
| Left | 53±8 % | 62±9 % | 59±3 % | 62±3 % | 65±7 % |
| Right | 47±8 % | 38±9 % | 41±3 % | 38±3 % | 35±7 % |
| Upper | 56±13 % | 43±5 % | 34±3 % | 33±4 % | 40±1 % |
| Lower | 44±13 % | 57±5 % | 66±3 % | 67±4 % | 60±1 % |

## 4.4 Discussion and Conclusion

To compare our findings with those in [74], we used a similar bandpass filter, with a lower cut-off frequency $f_L = 150$ Hz and an upper cut-off frequency $f_H = 250$ Hz (see Section 4.2.2). Although we lose important information from the signal in the higher frequency range, due to the dominance of the signal energy in the low frequency range, a higher cut-off frequency $f_H$ would not have a huge impact on the acoustic thoracic images. According to Figure 4.3 a bandpass filter with an upper cut-off frequency of $f_H \approx 600$ Hz could be considered.

Our findings regarding amplitude and regional distribution of lung sounds correspond most closely with those in [74], although we recorded the lung sounds in supine position. The authors in [78] already observed that, compared with sitting, the supine position does not cause a substantial change in lung sound intensity. The authors in [76] also observed a constant regional distribution for RATHIs at airflow rates of 1.0, 1.5 and 2.0 l/s. The authors in [74] showed the same for Vibration Response Images at airflow rates of 1.0, 1.3 and 1.7 l/s. Regarding the relationship between airflow rate and the square root of lung sound energy (see Section 4.3.1), the authors in [74] achieved for linear regression a coefficient of determination of $R^2 = 0.95$.

Limitations of our experiment are the small number of subjects and the lack of female subjects.

To conclude this chapter, we summarize our observations as follows. We observe a linear dependence between airflow rate and lung sound amplitude. In our recordings, the signal energy from lung sounds over the left hemithorax is distinctly higher than from those over the right one. Above an airflow rate of 0.7 l/s, we observe a constant regional distribution for the lung sound energy. Although we recorded lung sounds on the posterior chest in supine position instead of sitting, our findings match most closely with those in [74].

# 5 Clinical Trial - Multi-channel Lung Sound Database

We conducted a clinical trial for the recording of a multi-channel lung sound database. The clinical trial design is described in Section 5.1 and the considered lung diseases in Section 5.2. The resulting database, which we used for our experiments in Chapter 8 and Chapter 9, is presented in Section 5.3.

## 5.1 Clinical Trial

We performed a prospective monocentric medical device clinical trial. Details are presented in the following subsections.

### 5.1.1 Inclusion Criteria, Exclusion Criteria, Proband and Patient Enrollment

The minimum age of participants was 18 years and the gender ratio was balanced. We included non-smokers as well as smokers. Patients diagnosed with pneumothorax or IPF, and subjects from the lung-healthy control group were included. The control group was allowed to have a COPD lower or equal stage 2 according to the COPD GOLD criteria. Individuals with a body mass index (BMI)>30, a thorax surgery, or intubated and tracheostomy patients were excluded as prospective study candidates. Furthermore, in the control group, only individuals without respiratory system diseases and without prescriptions of medication influencing the respiration system were included. Pregnant and breastfeeding women were excluded.

Access to patients was granted by the *Division of Thoracic and Hyperbaric Surgery* and the *Division of Pulmonology* at the *Medical University of Graz (MUG)*.

### 5.1.2 Examination of a Subject

The examination workflow is illustrated in Figure 5.1. For the participation in the study, the subjects had to give written and oral consent. During a first screening they were assigned to one of the three groups[2], i.e. to *lung-healthy*, *pneumothorax*, or *IPF*. Then, subject data was gathered,

[2] Pathological subjects were diagnosed by a qualified physician before inclusion.

including age, gender, height, weight, and BMI. Following this, several medical investigations[3] were carried out, such as blood pressure reading, pulse measurement, oxygen saturation measurement, lung auscultation, and lung function testing. The last part of the screening was the anamnesis. If the subject was still eligible for further participation, we continued with the study specific interventions. This included the gathering of study data and the recording of lung sounds at different airflow values over 30 seconds, respectively.

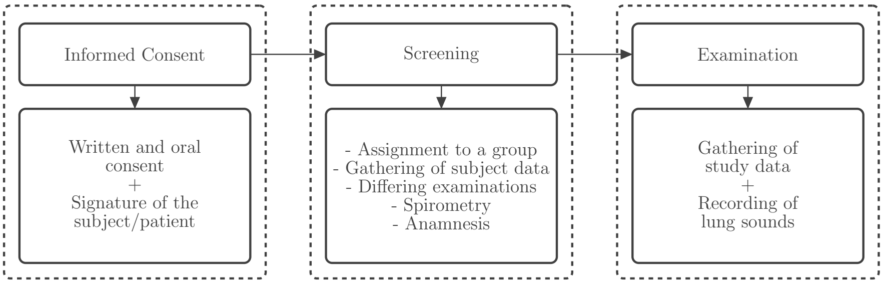

Informed Consent
Written and oral consent + Signature of the subject/patient
Screening
- Assignment to a group
- Gathering of subject data
- Differing examinations
- Spirometry
- Anamnesis
Examination
Gathering of study data + Recording of lung sounds

*Figure 5.1: Examination workflow for the clinical trial.*

### 5.1.3 Benefit and Risk Assessment

The risk associated with the clinical trial was very low. The measurement procedure caused extremely low inconvenience for the subjects. A physician was present during the entire examination. For a detailed assessment of the risks related to our MLSRD, we performed a risk analysis during the development process of the device. Furthermore, we implemented appropriate hardware improvements to keep potential risks in an acceptable range. In general, the risk-benefit balance can be assessed positively due to the implemented precautions and the low potential risk of complications associated with the recording procedure.

### 5.1.4 Statistical Methodology and Analysis

The task from a statistical point of view is the modeling of data. We want to distinguish between normal and abnormal lung sounds using statistical models. To successfully approach this task, we tried to collect adequate data. Several physiological features have a minor effect on lung sounds, such as gender, subcutaneous fat layer, and age [79–81]. Although these impacts are small, we considered the following groups:

- Uniform distribution of study participants over gender (*male*, *female*).
- Age is subdivided into three groups, i.e. 18 to 39 years, 40 to 59 years, and >60 years.
- Body mass index is categorised into three groups, i.e. BMI <20, 20 to <25, and 25 to <30.

[3] Within this thesis, we consider only the acoustic signals for classification, i.e. only the lung sounds. For this reason, we do not provide further details on the gathered data from the medical investigations.

The combination of all groups results in 18 categories. We aimed to include two subjects per category. Hence, the control group consists of 36 subjects. For the categories *pneumothorax* and *IPF*, we aimed to record 10 patients per disease, respectively. Due to the influence of airflow rate on the characteristics of lung sounds, we conducted several recordings at different airflow rate for each subject. In total, we aimed to record data from 56 individuals. The number of subjects included in the clinical trial is based on available resources and not on classical *sample size* considerations, i.e. it is a realistic estimate of the number of recruitable patients. From a machine learning perspective, the amount of data should be as large as possible to obtain well-performing models.

### 5.1.5 Documentation and Data Management

A screening and enrollment log was completed for all eligible and non-eligible subjects. A case report form (CRF) was filled out and signed by the principal investigator or co-investigator after inclusion of the subjects. All entries were checked by authorised personnel (i.e. Monitor).

### 5.1.6 Ethical and Legal Aspects

**Informed consent of subjects:** All subjects/patients had to give oral and written consent in order to participate in the study.
**Ethics and legal requirements:** We ensured that the clinical trial was conducted in full conformance with the principles of the *Declaration of Helsinki*. Furthermore, the following guidelines and laws were addressed:

- Austrian medical device act (as actual amended)
- Good clinical practice (EN ISO 14155)
- ICH-GCP guidelines
- Council Directive 93/42/EEC

**Acknowledgment / approval of the clinical investigation:** The clinical trial was approved by the Ethics Committee of the Medical University of Graz (Reference number: 28-088 ex 15/16).
**Insurance:** The subjects were insured during their participation in the clinical trial according to legal requirements.

## 5.2 Considered Lung Diseases

### 5.2.1 Idiopathic Pulmonary Fibrosis

In IPF, scars are formed in the lung tissue. Early diagnosis and a timely therapy is crucial for a successful treatment [82], i.e. any delay leads to a higher mortality rate [83]. According to [84], the incidence of IPF appears to be on the rise, and prevalence is expected to increase with aging population. Currently, CT scan is the gold standard for the diagnosis of pulmonary fibrosis.

In terms of auscultation, early markers are *velcro* crackles present in more than 80% of IPF patients [82,85]. In particular, inspiratory fine (or *velcro*) crackles are heard over affected areas [38, 86]. Crackles appear first in the basal areas and with further progression of the disease also in the upper areas of the lung [85]. The authors of [85] promote the assessment of *velcro* crackles by lung auscultation as the only realistic means for early detection. The recent success to slow down the disease progression renders this even more relevant.

### 5.2.2 Pneumothorax

A pneumothorax is a condition with an abnormal collection of air in the pleural space. The gold standard for diagnosis of pneumothorax is currently the CT scan [87]. This is often replaced by ultrasound. However, the diagnosis with lung X-ray carries the risk of an occult pneumothorax resulting in high uncertainty of this examination modality [88].

The presence of a pneumothorax has significant impact on the characteristics of lung sounds, since the resonance chamber within the hemithorax is altered [89]. Auscultatory findings are subdued or absent lung sounds over the affected area [40].

## 5.3 Multi-channel Lung Sound Database

Due to recruitment problems, we were not able to fill all the categories as intended according to Section 5.1.4. We included in total 24 subjects, i.e. 16 *lung-healthy* subjects, seven *IPF* patients, and one *pneumothorax* patient. Due to just a single *pneumothorax* subjects, we excluded this category from the final datasets. An overview on the clinical trial subjects is given in Table 5.1.

For the measurements, we defined the distance between the 7th cervical vertebra (C7) and the center line of the topmost row of sensors as $d \approx 10$ cm. Minor variations in favour of better sensor attachment were possible. For each subject, we included 16-channel lung sound recordings at two different maximum inspiratory airflow values. This results in two 16-channel lung sound recordings at varying airflow rates for each of the 23 subjects, with several breathing cycles within 30 seconds, respectively. During the measurements, we did not try to enforce having the same airflow values for each subject, because this could have resulted in discomfort for some of the pathological ones. Instead, the given maximum inspiratory airflow values were specified depending on the acceptability of the subjects, ranging from shallow to deep breathing. We picked the included measurements by having random airflow value above 0.5 l/s.

*Table 5.1: Subject and measurements contained in the multi-channel lung sound database [90].*

| Subject # | Gender | Age | Height | Weight | BMI | Category | Max. Insp. Airflow [l/s] |
|---|---|---|---|---|---|---|---|
| 1 | male | 27 | 178 | 78 | 24.6 | lung-healthy | 1.0 & 1.5 |
| 2 | male | 42 | 167 | 62 | 22.2 | lung-healthy | 1.0 & 1.3 |
| 3 | male | 26 | 189 | 75 | 21.0 | lung-healthy | 1.0 & 1.2 |
| 4 | male | 30 | 193 | 74 | 19.9 | lung-healthy | 1.2 & 1.5 |
| 5 | male | 27 | 173 | 85 | 28.4 | lung-healthy | 1.0 & 1.3 |
| 6 | male | 23 | 193 | 70 | 18.8 | lung-healthy | 0.6 & 1.0 |
| 7 | male | 41 | 180 | 97 | 29.9 | lung-healthy | 0.5 & 1.2 |
| 8 | male | 28 | 172 | 82 | 27.7 | lung-healthy | 0.5 & 1.0 |
| 9 | male | 53 | 180 | 80 | 24.7 | lung-healthy | 0.7 & 1.7 |
| 10 | male | 43 | 178 | 78 | 24.6 | lung-healthy | 1.5 & 2.0 |
| 11 | female | 30 | 166 | 58 | 21.0 | lung-healthy | 1.0 & 1.2 |
| 12 | female | 24 | 172 | 73 | 24.7 | lung-healthy | 0.7 & 1.3 |
| 13 | female | 27 | 172 | 56 | 18.9 | lung-healthy | 0.7 & 1.5 |
| 14 | female | 53 | 160 | 69 | 27.0 | lung-healthy | 0.5 & 0.7 |
| 15 | female | 30 | 160 | 50 | 19.5 | lung-healthy | 0.5 & 1.0 |
| 16 | female | 43 | 163 | 60 | 22.6 | lung-healthy | 1.2 & 1.5 |
| 17 | male | 76 | 184 | 92 | 27.2 | IPF | 0.8 & 1.0 |
| 18 | male | 60 | 175 | 82 | 26.8 | IPF | 1.0 & 2.0 |
| 19 | male | 79 | 175 | 75 | 24.5 | IPF | 1.0 & 1.2 |
| 20 | male | 74 | 187 | 83 | 23.7 | IPF | 1.0 & 1.2 |
| 21 | male | 71 | 178 | 80 | 25.2 | IPF | 0.7 & 0.8 |
| 22 | male | 74 | 169 | 74 | 25.9 | IPF | 0.7 & 1.0 |
| 23 | female | 76 | 158 | 53 | 21.2 | IPF | 0.5 & 1.0 |

# 6 Deep Neural Networks

In this chapter, we provide some basics for Chapter 7, 8, and 9. In Section 6.1, we introduce several deep neural network (DNN) architectures, such as MLPs, *vanilla* RNNs, long short-term memory (LSTM) networks, gated recurrent units (GRUs), bidirectional recurrent neural networks (BiRNNs), and CNNs. Furthermore, in Section 6.2, we discuss some regularization methods for neural networks, including virtual adversarial training (VAT), dropout, and noise injection.

*Some parts of this chapter were published in the IEEE Transactions on Biomedical Engineering in 2018 under the title of "Heart Sound Segmentation - An Event Detection Approach using Deep Recurrent Neural Networks" [91]. Furthermore, some parts are also contained in the IEEE Journal of Biomedical and Health Informatics (submitted 2019) under the title of "Multi-channel Lung Sound Classification with Convolutional Recurrent Neural Networks" [90]. As minor modifications, we changed some wordings, added some sentences, and updated the figures.*

## 6.1 Neural Network Architectures

### 6.1.1 Multilayer Perceptrons

MLPs [92] are the simplest type of artificial neural networks. In an MLP, information flows forward through the network, i.e. the output of the model is not fed back into itself. Equations (6.1-6.2) describe the MLP mathematically.

$$\mathbf{h}_f^l = g(\mathbf{W}_x^l \mathbf{x}_f^l + \mathbf{b}_h^l) \tag{6.1}$$

$$\mathbf{y}_f = m(\mathbf{W}_y \mathbf{h}_f^{L-1} + \mathbf{b}_y) \tag{6.2}$$

It consists of several layers $L$, with $l \in \{1, ..., L-1\}$ being the index of the hidden layers. For a frame-wise processing, $f \in \{1, ..., F\}$ indicates the frame index, i.e. the processing at a certain time step, with $F$ being the number of frames. In a first step, the dot product $\mathbf{W}_x^l \mathbf{x}_f^l$ and the bias term $\mathbf{b}_h^l$ are summed up, with $\mathbf{x}_f^l$ being the input vector and $\mathbf{W}_x^l$ the input weight matrix. Following this, a non-linear function $g(\cdot)$ is applied to obtain the hidden states $\mathbf{h}_f^l$, which are used as input for the next hidden layer $\mathbf{x}_f^{l+1}$. The states of the last hidden layer $\mathbf{h}_f^{L-1}$ are fed into the output layer. According to Equation (6.2), the sum between the dot product $\mathbf{W}_y \mathbf{h}_f^{L-1}$ and the bias term

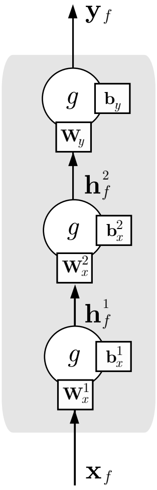


*Figure 6.1: Flow graph of an MLP with two hidden layers.*

$\mathbf{b}_y$ is computed, with $\mathbf{W}_y$ being the output weight matrix. With a non-linear function $m(\cdot)$ the output $\mathbf{y}_f$ is obtained. A deep MLP is formed by stacking several hidden layers. The network is trained with back-propagation by using a differentiable cost function.

MLPs can be used to process sequential input, by classifying successive short time frames of the sequence. In the simplest case, classification of the single frames happens independently from each other. One exemplary approach to incorporate information of neighbouring frames would be an appropriate feature extraction.

### 6.1.2 Recurrent Neural Networks

More suitable models to process sequential input and to learn long temporal dependencies within the data are RNNs [93]. Several RNN architectures exist, including *vanilla* RNNs, LSTM networks [94, 95], and gated recurrent neural network (GRNN) [96, 97]. In contrast to *vanilla* RNNs, LSTMs and GRNNs can model longer temporal dependencies. GRNNs are simplifications of LSTMs, which achieve comparable performance with less parameters.

#### Vanilla Recurrent Neural Network

The flow graph of a *vanilla* RNN hidden layer is shown in Figure 6.2. The output of a hidden recurrent layers $\mathbf{h}_f^l$ is computed as

$$\mathbf{h}_f^l = g(\mathbf{W}_x^l \mathbf{x}_f^l + \mathbf{W}_h^l \mathbf{h}_{f-1}^l + \mathbf{b}_h^l). \tag{6.3}$$

First of all, the dot product $\mathbf{W}_x^l \mathbf{x}_f^l$, the projected previous hidden state $\mathbf{W}_h^l \mathbf{h}_{f-1}^l$, and the bias term $\mathbf{b}_h^l$ are summed up. $\mathbf{x}_f^l$ is the input vector, $\mathbf{h}_{f-1}^l$ the previous recurrent hidden state vector, $\mathbf{W}_x^l$ the input weight matrix, and $\mathbf{W}_h^l$ the hidden weight matrix. The output of the hidden layer

$\mathbf{h}_f^l$ is obtained with a non-linear function $g(\cdot)$. The output of the RNN is computed according to Equation (6.2).

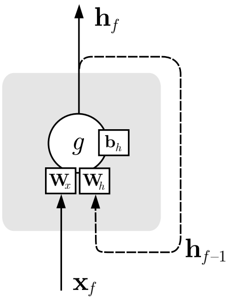


*Figure 6.2: Flow graph of a vanilla RNN hidden layer.*

### Long Short Term Memory Networks

LSTMs [94, 95] are temporal recurrent neural networks using memory cells to store temporal information. In contrast to RNNs, LSTMs have memory cells, which store or erase their content using *input* gates $i$ or *forget* gates $u$. An additional *output* gate $o$ is used to access this information. Figure 6.3 shows the flow-graph of an LSTM layer.

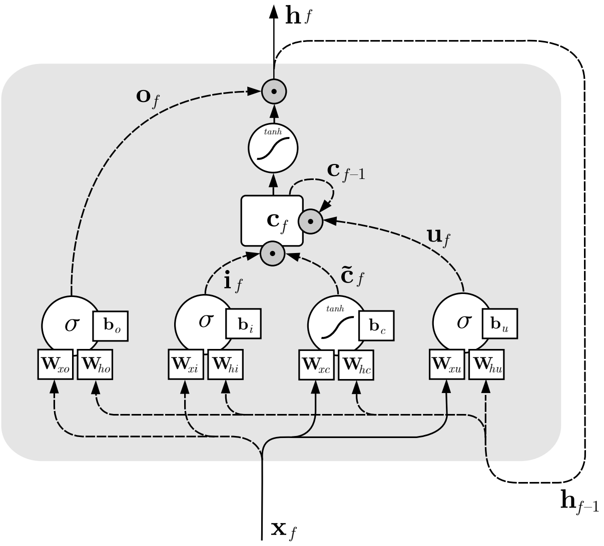


*Figure 6.3: Flow graph of a LSTM hidden layer [98]. $\boldsymbol{i}_f$, $\boldsymbol{u}_f$, and $\boldsymbol{o}_f$ are the input, forget and output states, respectively. $\mathbf{c}_f$ denote the memory cell and $\tilde{\mathbf{c}}_f$ the new memory cell content.*

In Equations (6.4-6.9), the network is described mathematically. The input states $\mathbf{i}_f^l$ are calculated by applying a sigmoid function $\sigma$ to the sum of the dot-product of the input weight matrix $\mathbf{W}_{xi}^l$ and the inputs $\mathbf{x}_f^l$, the projected previous hidden states $\mathbf{W}_{hi}^l\mathbf{h}_{f-1}^l$ and the bias vector $\mathbf{b}_i^l$ of layer $l$ (cf. Equation 6.4). The forget states $\mathbf{u}_f^l$ (cf. Equation 6.5) and output states $\mathbf{o}_f^l$ (cf. Equation 6.6) are computed in a similar way, except using individual *forget* matrices $\mathbf{W}_{xu}^l, \mathbf{W}_{hu}^l$ and the *forget* bias vector $\mathbf{b}_u^l$ and *output* matrices $\mathbf{W}_{xo}^l, \mathbf{W}_{ho}^l$ and the *output* bias vector $\mathbf{b}_o^l$, respectively.

The new memory states $\tilde{\mathbf{c}}_f^l$ are obtained by applying a *tanh* activation function on the sum of the projected inputs $\mathbf{W}_{xc}^l\mathbf{x}_f^l$, previous hidden memory states $\mathbf{W}_{hc}^l\mathbf{h}_{f-1}^l$ and the bias vector $\mathbf{b}_c^l$ in Equation (6.7). The memory cell states $\mathbf{c}_f^l$ are updated by the previous memory states $\mathbf{c}_{f-1}^l$ and $\tilde{\mathbf{c}}_f^l$ (cf. Equation (6.8)), weighted by the forget states $\mathbf{u}_f^l$ and the input state $\mathbf{i}_f^l$, respectively ($\odot$ denotes an element-wise product). The outputs $\mathbf{h}_f^l$ are computed with the current memory states $\tanh(\mathbf{c}_f^l)$ and the output states $\mathbf{o}_f^l$ in Equation (6.9).

$$\mathbf{i}_f^l = \sigma(\mathbf{W}_{xi}^l\mathbf{x}_f^l + \mathbf{W}_{hi}^l\mathbf{h}_{f-1}^l + \mathbf{b}_i^l) \tag{6.4}$$

$$\mathbf{u}_f^l = \sigma(\mathbf{W}_{xu}^l\mathbf{x}_f^l + \mathbf{W}_{hu}^l\mathbf{h}_{f-1}^l + \mathbf{b}_u^l) \tag{6.5}$$

$$\mathbf{o}_f^l = \sigma(\mathbf{W}_{xo}^l\mathbf{x}_f^l + \mathbf{W}_{ho}^l\mathbf{h}_{f-1}^l + \mathbf{b}_o^l) \tag{6.6}$$

$$\tilde{\mathbf{c}}_f^l = \tanh(\mathbf{W}_{xc}^l\mathbf{x}_f^l + \mathbf{W}_{hc}^l\mathbf{h}_{f-1}^l + \mathbf{b}_c^l) \tag{6.7}$$

$$\mathbf{c}_f^l = \mathbf{u}_f^l \odot \mathbf{c}_{f-1}^l + \mathbf{i}_f^l \odot \tilde{\mathbf{c}}_f^l \tag{6.8}$$

$$\mathbf{h}_f^l = \mathbf{o}_f^l \odot \tanh(\mathbf{c}_f^l) \tag{6.9}$$

In classical RNNs, the hidden activation is recomputed at each time-step (cf. Equation 6.3). LSTMs are able to decide whether to keep or erase existing information with the help of their gates. If LSTMs detect important features from an input sequence at early stage, they easily carry this information over a long distance in time, hence, potentially capturing long-distance dependencies.

### Gated Recurrent Unit

Gating mechanism in recurrent neural networks are GRUs [96, 97]. They have *reset*- and *update*-gates, which are coupling static and temporal information. The flow graph for GRNN hidden layer is shown in Figure 6.4. Equations (6.10-6.13) describe a GRNN hidden layer mathematically:

$$\mathbf{h}_f^l = (1 - \mathbf{z}_f^l) \odot \mathbf{h}_{f-1}^l + \mathbf{z}_f^l \odot \tilde{\mathbf{h}}_f^l \tag{6.10}$$

$$\mathbf{z}_f^l = \sigma(\mathbf{W}_{xz}^l\mathbf{x}_f^l + \mathbf{W}_{hz}^l\mathbf{h}_{f-1}^l + \mathbf{b}_z^l) \tag{6.11}$$

$$\tilde{\mathbf{h}}_f^l = g(\mathbf{W}_{xh}^l\mathbf{x}_f^l + \mathbf{W}_{hh}^l(\mathbf{r}_f^l \odot \mathbf{h}_{f-1}^l) + \mathbf{b}_h^l) \tag{6.12}$$

$$\mathbf{r}_f^l = \sigma(\mathbf{W}_{xr}^l\mathbf{x}_f^l + \mathbf{W}_{hr}^l\mathbf{h}_{f-1}^l + \mathbf{b}_r^l) \tag{6.13}$$

In Equation (6.10), the output states $\mathbf{h}_f^l$ are computed by linearly interpolating between past states $\mathbf{h}_{f-1}^l$ and current information $\tilde{\mathbf{h}}_f^l$, using the *update*-states $\mathbf{z}_f^l$ ($\odot$ denotes an element-wise product). The *update*-states $\mathbf{z}_f^l$ are computed in Equation (6.11) as a sigmoid function of the weighted input $\mathbf{x}_f^l$ and the past hidden states $\mathbf{h}_{f-1}^l$, with weights $\mathbf{W}$ and bias term $\mathbf{b}$. The *update*-gates $z$ decide to renew the current state of the model. According to Equation 6.12, the states $\tilde{\mathbf{h}}_f^l$ are computed by applying a non-linear function $g(\cdot)$ to the input and the previous hidden states

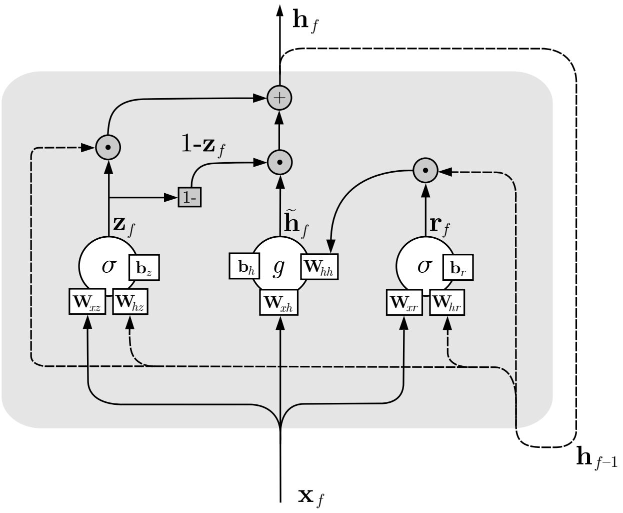


*Figure 6.4: Flow graph of a GRNN hidden layer [91, 98]. The reset and update states are $\boldsymbol{r}_f$ and $\boldsymbol{z}_f$, and $\boldsymbol{h}_f$ and $\tilde{\boldsymbol{h}}_f$ denote the activation and the candidate activation, respectively.*

$\mathbf{h}^l_{f-1}$. The reset state $\mathbf{r}^l_f$ is computed with a sigmoid function of the current input $\mathbf{x}^l_f$ and the past states $\mathbf{h}^l_{f-1}$ (cf. Equation (6.13)). It enables to delete the current state of the model, allowing to forget the previously computed information. The output of the GRNN is computed according to Equation (6.2).

### Bidirectional Recurrent Neural Networks

Extensions of conventional RNNs are their bidirectional implementations [99]. In addition to past information, BiRNNs exploit future information as well. This is achieved by processing data in both directions with two separate hidden layers (see Figure 6.5). BiRNNs combined with GRNNs are bidirectional gated recurrent neural networks (BiGRNNs) [30]. Due to simplicity, we explain the extension to a bidirectional network just for the *vanilla* RNN.

Equations (6.14-6.16) describe a BiGRNN mathematically.

$$\overrightarrow{\mathbf{h}}^l_f = g(\mathbf{W}^l_{x\overrightarrow{h}}\mathbf{x}^l_f + \mathbf{W}^l_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{\mathbf{h}}^l_{f-1} + \mathbf{b}^l_{\overrightarrow{h}}) \tag{6.14}$$

$$\overleftarrow{\mathbf{h}}^l_f = g(\mathbf{W}^l_{x\overleftarrow{h}}\mathbf{x}^l_f + \mathbf{W}^l_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{\mathbf{h}}^l_{f-1} + \mathbf{b}^l_{\overleftarrow{h}}) \tag{6.15}$$

$$\mathbf{y}_f = m(\mathbf{W}_{\overrightarrow{h}y}\overrightarrow{\mathbf{h}}^{L-1}_f + \mathbf{W}_{\overleftarrow{h}y}\overleftarrow{\mathbf{h}}^{L-1}_f + \mathbf{b}_y) \tag{6.16}$$

$g(\cdot)$ and $m(\cdot)$ are non-linear functions, and $\mathbf{W}$ and $\mathbf{b}$ denote the weights and bias terms. The forward hidden sequence $\overrightarrow{\mathbf{h}}^l_f$ is generated by processing the data in the forward layer from $f = 1$ to $F$. The backward hidden sequence $\overleftarrow{\mathbf{h}}^l_f$ is generated by processing the data in the backward layer from $f = F$ to 1. Both hidden sequences $\overrightarrow{\mathbf{h}}^l_f$ and $\overleftarrow{\mathbf{h}}^l_f$ are fed to the next hidden layer $l + 1$. The hidden activations $\overrightarrow{\mathbf{h}}^{L-1}_f$ and $\overleftarrow{\mathbf{h}}^{L-1}_f$ of the last hidden layer $L - 1$ are fed to the output layer (cf. Equation (6.16)).

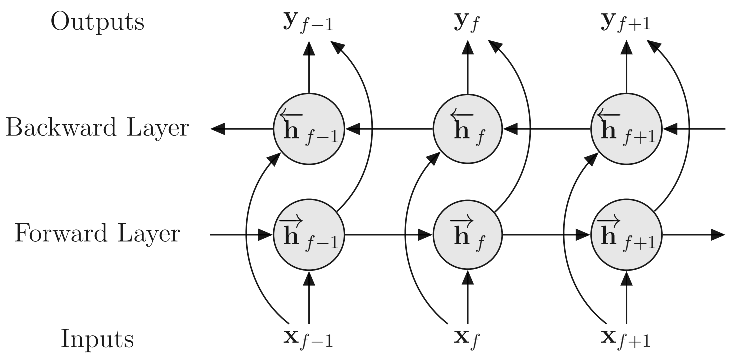

Outputs
Backward Layer
Forward Layer
Inputs
$\mathbf{y}_{f-1}$
$\mathbf{y}_f$
$\mathbf{y}_{f+1}$
$\overleftarrow{\mathbf{h}}_{f-1}$
$\overleftarrow{\mathbf{h}}_f$
$\overleftarrow{\mathbf{h}}_{f+1}$
$\overrightarrow{\mathbf{h}}_{f-1}$
$\overrightarrow{\mathbf{h}}_f$
$\overrightarrow{\mathbf{h}}_{f+1}$
$\mathbf{x}_{f-1}$
$\mathbf{x}_f$
$\mathbf{x}_{f+1}$

*Figure 6.5: Bidirectional recurrent neural network [30].*

### 6.1.3 Convolutional Neural Networks

CNNs are feedforward neural networks, which are used to process data that has a known grid-like topology, such as time series (1-dimensional) and image data (2-dimensional) [92]. They are widely applied in computer vision and audio processing [100].

CNNs consist of three different types of layers, i.e. convolutional layers, pooling layers, and fully-connected layers. Figure 6.6 illustrates the convolutional layer and the pooling layer conceptually.

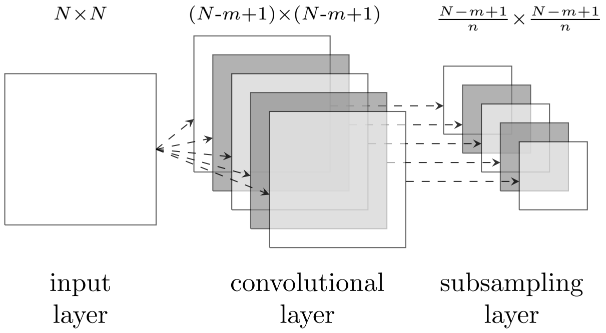

$N \times N$
$(N\text{-}m\text{+}1) \times (N\text{-}m\text{+}1)$
$\frac{N-m+1}{n} \times \frac{N-m+1}{n}$
input layer
convolutional layer
subsampling layer

*Figure 6.6: Schematic illustration of a convolutional layer (with $K = 5$ feature maps) and a subsampling (pooling) layer.*

We consider an $N \times N$ input layer (e.g. a grayscale image), followed by a convolutional layer. The convolutional layer performs an image convolution of the input layer. An $m \times m$ kernel (filter), with a stride of one, results in a feature map (i.e. convolutional layer output) with the size $(N - m + 1) \times (N - m + 1)$. Zero padding beyond the borders of the input image can be used to retain the original shape $N \times N$. Each convolutional layer consists of several feature maps $K$. Equation 6.17 describes the operations to obtain the feature map $k \in \{1, ..., K\}$ in layer $l$

mathematically.

$$h_{ij}^{kl} = g(\mathbf{W}^{kl} * \mathbf{X}_{ij}^{l} + b_k^l) \tag{6.17}$$

A section of the input image $\mathbf{X}^l$, with size $m \times m$ and position $i, j \in \{1, ..., N-m+1\}$, is convolved with a kernel $\mathbf{W}^{kl}$, and a bias term $b_k^l$ is added. The activation $h_{ij}^{kl}$ is obtained after applying a non-linear function $g$.

The convolutional layer can be followed by a subsampling layer, which replaces the output of the convolutional layer at a certain location with a summary statistic of the nearby outputs. Different pooling methods exist, such as *max pooling*, *average pooling*, and $L^2$ *norm pooling*. In *max pooling* the maximum output within a rectangular $n \times n$ neighborhood is obtained. Given a previous layer with dimensions $(N-m+1) \times (N-m+1)$, the output of a max pooling layer is a $\frac{N-m+1}{n} \times \frac{N-m+1}{n}$ layer. With pooling, the representations get approximately invariant to small translations of the input [92].

After several convolutional and subsampling layers a fully-connected layer is used.

## 6.2 Regularization

Deep neural networks usually require many training samples. To prevent overfitting on small datasets, we consider three different approaches for regularization to improve the ability of the models to generalize to test data, i.e. virtual adversarial training, dropout, and noise injection.

### 6.2.1 Virtual Adversarial Training

VAT [101,102] is a regularization method that makes the model robust against adversarial perturbations [103,104]. It promotes local smoothness of the posterior distribution $p(\mathbf{y}_f|\mathbf{x}_f)$ with respect to $\mathbf{x}_f$. The posterior distribution, or more precisely the softmax activation of the network output $\mathbf{h}_f^l$, should vary minimally for small, bounded perturbations of the input $\mathbf{x}_f$. The adversarial perturbation $\boldsymbol{\delta}_f$ is determined on frame-level by maximizing the Kullback–Leibler (KL)-divergence $(\cdot||\cdot)$ of the posterior distribution for unperturbed and perturbed inputs, i.e.

$$\boldsymbol{\delta}_f = \underset{||\boldsymbol{\delta}||<\epsilon}{\arg\max}\, \mathrm{KL}(p(\mathbf{y}|\mathbf{x}_f)||p(\mathbf{y}|\mathbf{x}_f + \boldsymbol{\delta}), \tag{6.18}$$

where $\epsilon > 0$ limits the maximum perturbation, i.e. the noisy input $\mathbf{x}_f + \boldsymbol{\delta}$ lies within a radius $\epsilon$ around $\mathbf{x}_f$. The smaller the $\mathrm{KL}(p(\mathbf{y}|\mathbf{x}_f)||p(\mathbf{y}|\mathbf{x}_f + \boldsymbol{\delta}_f)$, the smoother the posterior distribution is around $\mathbf{x}_f$. Instead of maximizing the conditional likelihood $p(\mathbf{y}_f|\mathbf{x}_f)$ of the model during training, we maximize the regularized objective

$$\sum_f \log\big(p(\mathbf{y}_f|\mathbf{x}_f)\big) - \lambda \sum_f \mathrm{KL}(p(\mathbf{y}|\mathbf{x}_f)||p(\mathbf{y}|\mathbf{x}_f + \boldsymbol{\delta}_f)), \tag{6.19}$$

where the tradeoff parameter $\lambda$ and the radius $\epsilon$ have to be selected on a validation set. For further details regarding the implementation, we refer to [101]. Tunable parameters are the number of iterations $I_p$, the radius $\epsilon$, and the tradeoff parameter $\lambda$.

### 6.2.2 Dropout

The idea of dropout is to randomly drop units from the neural network during training [105]. In this work, we consider input dropout applied on the hidden layers. Due to simplicity, we show dropout just for the *vanilla* RNN. Equations (6.20-6.22) describe the feed-forward operation of the network with dropout.

$$\mathbf{v}^l \sim Bernoulli(\mathbf{p}) \tag{6.20}$$

$$\tilde{\mathbf{x}}_f^l = \mathbf{v}^l \odot \mathbf{x}_f^l \tag{6.21}$$

$$\mathbf{h}_f^l = g(\mathbf{W}_x^l \tilde{\mathbf{x}}_f^l + \mathbf{W}_h^l \mathbf{h}_{f-1}^l + \mathbf{b}^l) \tag{6.22}$$

For any hidden layer $l \in \{1, ..., L-1\}$, $\mathbf{v}^l$ is a vector of independent Bernoulli random variables, each having a probability $p$ of being 1, with $\mathbf{p} = [p, p, ..., p]^T$. The vector $\mathbf{v}^l$ is multiplied element-wise with the inputs of the layer $\mathbf{x}_f^l$, to create the thinned inputs $\tilde{\mathbf{x}}_f^l$. The thinned inputs are then used as inputs to the current layer. For training, the derivatives of the loss function are backpropagated through the sub-network. For testing, the network is used without dropout and the weights are scaled as $\mathbf{W}_{x,test}^l = p\mathbf{W}_x^l$.

### 6.2.3 Noise Injection

Noise injection to the inputs of a neural network can be considered as a form of data augmentation [92]. The authors in [106] showed that noise injection can be very effective if the noise magnitude is carefully tuned. Dropout (see Section 6.2.2) can be considered as a process of constructing new inputs by using a particular type of noise [92]. We add zero mean Gaussian noise to the inputs $\mathbf{x}_f$ and the hidden units during training. Standard deviation and noise level are tuned.

# 7 Excursion to Heart Sound Segmentation

Due to a large public available heart sound dataset, including appropriate labeling of events, we made an excursion to heart sound segmentation to start the development of an event detection framework. The idea is to apply the framework for event detection in lung sounds afterwards (see Section 8). In the context of lung sound analysis, we consider heart sounds as noise. In our MLSRD, we use a high-pass filter to remove the main components of disturbing heart sounds (cf. Section 3.3.1). The dominant frequency range of lung sounds is between 150 Hz and 2 kHz and that of heart sounds is below 150 Hz [4].

This chapter is organised as follows. After a short introduction in Section 7.1, we present the proposed processing framework for heart sound segmentation in Section 7.2. We describe the heart sound dataset in Section 7.3 and the feature extraction in Section 7.4. Before showing our experiments and results in Section 7.6, we describe the evaluation metrics in Section 7.5. We discuss the results and conclude this chapter in Section 7.7.

*This chapter was published in the IEEE Transactions on Biomedical Engineering in 2018 under the title "Heart Sound Segmentation - An Event Detection Approach using Deep Recurrent Neural Networks" [91]. We made minor modifications regarding some wordings and figures.*

## 7.1 Introduction

Computer-aided heart sound analysis can be considered as a twofold task: segmentation and subsequent classification. The accurate segmentation of the fundamental heart sounds, or more precisely of the state-sequence *first heart sound (S1) - systole - second heart sound (S2) - diastole*, is a challenging task. In heart sound recordings of healthy adults, only S1 and S2 are present. However, extra heart sounds (S3 and fourth heart sound (S4)) can occur during diastole, i.e. in the interval S2-S1, and heart murmurs during systole, i.e. in the interval (S1-S2), and/or diastole, as shown in Figure 7.1. Furthermore, the corruption by different noise sources (e.g. motion artefacts, ambient noise) and other body sounds (e.g. lung sounds, cough sounds) renders the segmentation even more challenging. According to [107], existing heart sound segmentation methods are classified into four groups: envelope-based methods [108–113], feature based methods [114–120], machine learning methods [121–127], and HMM methods[4] [128–134]. The authors in [134] introduced a

[4] We would rather include HMM methods in the machine learning category.

logistic regression hidden semi-Markov model (LR-HSMM) to predict the most likely sequence of states in the order of *S1 - systole - S2 - diastole*, using a priori information about expected durations of the heart sound states. In experiments, they achieve an average F-score of $F_1 = 95.63\,\%$ on an independent test set. Due to the significant improvement in comparison to other reported methods in the literature, it is considered as the state-of-the-art method by the authors in [107]. A more extensive evaluation of the algorithm on the 2016 PhysioNet/CinC Challenge data [107] is presented in [135]. The authors report an average F-score of $F_1 = 98.5\,\%$ for segmenting S1 and systole intervals and $F_1 = 97.2\,\%$ for segmenting S2 and diastole intervals. They observe detection errors especially in the situations of long heart cycles and irregular sinus rhythm. Also, the authors in [127] point out that the LR-HSMM-method [134] may be unsuitable for the segmentation in recordings with cardiac arrhythmia. Their main objective is to investigate if S1 and S2 can be detected without using a priori information about the state duration. They propose a machine learning approach with an MLP in combination with mel frequency cepstral coefficients (MFCCs)-features for S1 and S2 heart sound recognition. Using the K-means algorithm, they cluster the MFCC features into two groups to refine their representation and discriminative capability. The refined features are then fed to an MLP. In experiments with a relatively small dataset, the authors show that S1 and S2 can be detected with an accuracy of 91 %, outperforming well-known classifiers such as k-NN, GMMs, logistic regression, and SVMs.

Within this chapter, we exploit spectral information and temporal dependencies of heart sounds for heart sound segmentation. To this end, we propose an acoustic event detection approach with deep recurrent neural networks (DRNNs) [97,98,136]. RNNs are suitable to process sequential input of variable length, and learn temporal dependencies within the data [30, 137]. They are already used for heart sound *classification* [138–140], but to the best of our knowledge, not specifically for heart sound *segmentation*. Compared to the LR-HSMM-method [134], we do not directly

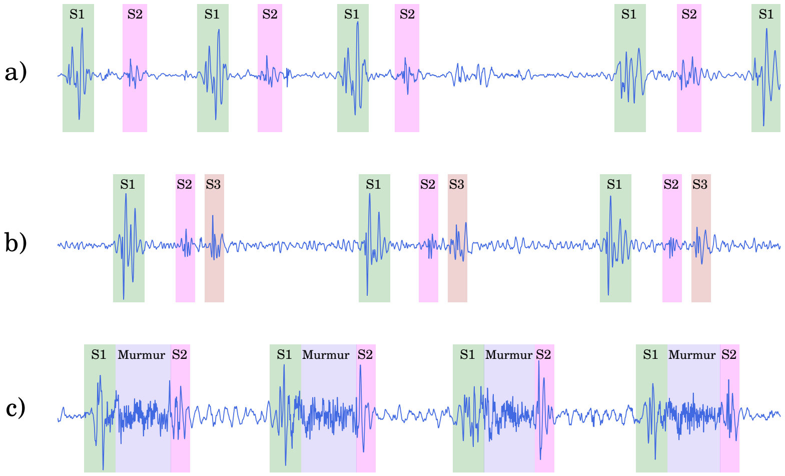


*Figure 7.1: Examples of heart sound recordings: a) cardiac arrhythmia, b) extra (third) heart sound S3, and c) heart murmur (mitral valve prolapse). The marked events are first (S1), second (S2) and third (S3) heart sounds and heart murmurs.*

incorporate a priori information about the state durations, because the model is capable of learning the temporal dependencies itself. Furthermore, we are flexible regarding the order of occurring states enabling to additionally model S3, S4 and heart murmurs. In acoustic event detection, sound events are usually detected by onsets and offsets, defining the beginning and ending of a particular event within an audio recording. It is differentiated between *polyphonic* and *monophonic* event scenarios: In the first case, multiple events can occur at the same time, whereas in the second case no overlapping events exist. Within this work, we consider heart sound segmentation as a *monophonic* event scenario, although heart sound recordings can be contaminated with body sounds and different noise sources, and therefore represent a *polyphonic* event scenario. DNNs show a significant boost in performance when applied to acoustic event detection. In particular, Gencoglu et al. [141] proposed an MLP architecture for acoustic event detection. Although MLPs are powerful network architectures, they do not model temporal context explicitly. To account for temporal structure, LSTM networks have been applied to acoustic keyword spotting [142] and polyphonic sound event detection [143]. LSTM networks are DNNs capable of modeling temporal dependencies. Performance in recognition comes at the expense of computational complexity and the amount of labeled data, required for training. LSTMs have a relatively high model complexity and parameter tuning is not always simple. A simplification of LSTMs are GRNNs, which have fewer parameters, but achieve comparable performance. Due to this fact, we focus on GRNNs for the accurate segmentation of fundamental heart sounds, or more precisely of the state-sequence *S1 - systole - S2 - diastole.* GRNNs already show promising results for acoustic event detection [31]. To exploit future information as well, and not just information from the past, we also consider bidirectional recurrent neural networks [99].

In particular, we extract spectral and envelope features from heart sounds and investigate the performance of different DRNN architectures to detect the state-sequence, i.e. acoustic events. We use data from the 2016 PhysioNet/CinC Challenge [107], containing heart sound recordings and annotations of the heart sound states. Our main contributions and results are:

- We compare different recurrent neural network architectures.
- We evaluate BiGRNNs in combination with VAT, dropout and data augmentation for regularization.
- We show state-of-the-art performance on the 2016 PhysioNet/CinC Challenge dataset.

## 7.2 Audio Processing Framework

Figure 7.2 shows the basic steps of our heart sound segmentation framework. Given the raw audio data $\mathbf{x}_t = [x_1, \dots, x_T]$, we extract a sequence of feature frames $\mathbf{x}_f \in \mathbb{R}^D$, where $f \in \{1, ..., F\}$

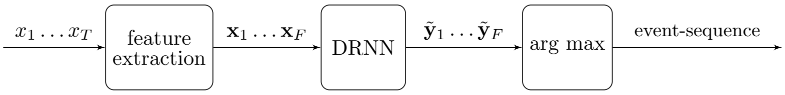

$x_1 \dots x_T$
feature extraction
$\mathbf{x}_1 \dots \mathbf{x}_F$
DRNN
$\tilde{\mathbf{y}}_1 \dots \tilde{\mathbf{y}}_F$
arg max
event-sequence

*Figure 7.2: Heart sound segmentation framework.*

is the frame index, with $F$ indicating the number of frames. These feature frames are processed by a multi-label DRNN with a softmax output layer. The index of the maximum value (*arg max*) of the real-valued output vector $\tilde{\mathbf{y}}_f$ determines the event class per frame, resulting in a sequence of frame labels as output. Consecutive identical frame labels are grouped as one event. A schematic illustration of the frame-wise single-channel lung sound processing framework is shown in Figure 7.3.

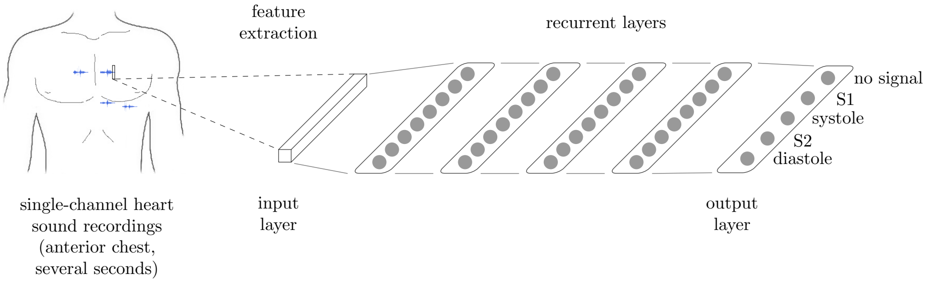


*Figure 7.3: Frame-wise single-channel heart sound processing framework with a recurrent neural network.*

# 7.3 Material - Heart Sound Database

## 7.3.1 Heart Sound Database

For the experiments within this section, we use heart sounds from the 2016 PhysioNet/CinC Challenge [107]. The dataset is a collection of several heart sound databases from different research groups, obtained in different real-world clinical and nonclinical environments. It contains recordings from normal subjects and pathological patients, which are grouped as follows: Normal control group (Normal), murmurs related to mitral valve prolapse (MVP), innocent or benign murmurs (Benign), aortic disease (AD), miscellaneous pathological conditions (MPCs), coronary artery disease (CAD), mitral regurgitation (MR), aortic stenosis (AS), and pathological (Pathologic). The heart sounds were recorded at the four common recording locations: aortic area, pulmonic area, tricuspid area, and mitral area. Due to the fact that the database is a collection of several small databases from different research groups, the recordings vary regarding several aspects: recording hardware, recording locations, data quality, and patient types as well as methods for identifying gold standard diagnoses. For further details, we refer to [107].

The training set includes data from six databases, with a total of 3153 heart sound recordings from 764 subjects/patients (see Table 7.1). The recordings are sampled with $f_s = 2\,\mathrm{kHz}$ and vary in length between 5 s and just over 120 s. The dataset is unbalanced, i.e. the number of normal recordings differ from that of abnormal recordings. Besides a binary diagnosis (-1=normal, 1=abnormal) for each heart sound recording, the challenge dataset further provides annotations for the heart sound states (*S1*, *systole*, *S2*, *diastole*). The annotations were generated with the LR-HSMM-based segmentation algorithm [134] (trained on PhysioNet (PN)-training-a) and further manually

*Table 7.1: Summary of the dataset (Training data from the 2016 PhysioNet/CinC Challenge) [107].*

| Challenge set | # Patients | # Recordings | # Beats |
|---|---|---|---|
| PN-training-a | 121 | 409 | 14559 |
| PN-training-b | 106 | 490 | 3353 |
| PN-training-c | 31 | 31 | 1808 |
| PN-training-d | 38 | 55 | 853 |
| PN-training-e | 356 | 2054 | 59593 |
| PN-training-f | 112 | 114 | 4260 |
| Total | 764 | 3153 | 84426 |

corrected. The annotations solely generated with the *segmentation algorithm* and those generated with the *segmentation algorithm and subsequent hand correction*, are accessible separately. In total 84426 beats were annotated in the PN-training set (after hand correction).

Because the reference annotations for the four heart sound states were not available for heart sound recordings marked with *unsure* (=low signal quality), we excluded these recordings. We further excluded areas labeled as *noisy* (labels: *'(N', 'N)'*) by setting the respective areas of the signal to zero (*no signal*). Table 7.2 shows the resulting number of recordings and beats.

*Table 7.2: Summary of the dataset (Training data from the 2016 PhysioNet/CinC Challenge) [107] after excluding areas labeled as noisy (labels: '(N', 'N)') and files marked as unsure.*

| Challenge set | # Recordings | # Beats |
|---|---|---|
| PN-training-a | 392 | 14559 |
| PN-training-b | 368 | 3353 |
| PN-training-c | 27 | 1808 |
| PN-training-d | 52 | 853 |
| PN-training-e | 1926 | 59567 |
| PN-training-f | 109 | 4260 |
| Total | 2874 | 84400 |

### 7.3.2 Training, Validation, and Test Data

Due to the fact that the original test set from the PhysioNet/CinC Challenge 2016 is not publicly available so far, we generated a new test-, validation- and training-set out of the original PhysioNet (PN)-training set (see Section 7.3.1). In the test set, we put exclusively PN-training-a and some recordings from PN-training-b and PN-training-e. For the recordings from PN-training-b and PN-training-e, we ensured their exclusivity in terms of subject affiliation, i.e. each subject is either only in the training set or the test set. We selected all recordings from the same subject with increasing 'Subject ID' (for PN-training-b) and increasing 'Raw record' name (for PN-training-e). This additional information is provided by the online appendix of the database. The resulting test set contains 764 recordings with 21116 beats in total. From the remaining recordings, we randomly selected 210 recordings (6135 beats) for the validation set and 1900 recordings (57149 beats) for

*Table 7.3: Summary of the test, validation and training set. The assigned number of recordings (#R.) and beats (#B.) are reported. The recordings are grouped as follows: Normal control group (Normal), murmurs related to MVP, innocent or benign murmurs (Benign), AD, MPCs, CAD, MR, AS, and pathological (Pathologic).*

| Dataset | Challenge set | Normal | | MVP | | Benign | | AD | | MPC | | CAD | | MR | | AS | | Pathologic | | **Total** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #R. | #B. | #R. | #B. | #R. | #B. | #R. | #B. | #R. | #B. | #R. | #B. | #R. | #B. | #R. | #B. | #R. | #B. | #R. | #B. |
| Test | PN-training-a | 116 | 4419 | 126 | 4583 | 114 | 4200 | 13 | 425 | 23 | 932 | - | - | - | - | - | - | - | - | 392 | 14559 |
| | PN-training-b | 135 | 1278 | - | - | - | - | - | - | - | - | 29 | 238 | - | - | - | - | - | - | 164 | 1516 |
| | PN-training-e | 205 | 5018 | - | - | - | - | - | - | - | - | 3 | 23 | - | - | - | - | - | - | 208 | 5041 |
| | Total | 456 | 10715 | 126 | 4583 | 114 | 4200 | 13 | 425 | 23 | 932 | 32 | 261 | - | - | - | - | - | - | **764** | **21116** |
| Validation | PN-training-b | 12 | 100 | - | - | - | - | - | - | - | - | 5 | 49 | - | - | - | - | - | - | 17 | 149 |
| | PN-training-c | 1 | 23 | - | - | - | - | - | - | - | - | - | - | 3 | 125 | 1 | 18 | - | - | 5 | 166 |
| | PN-training-d | 3 | 32 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 26 | 5 | 58 |
| | PN-training-e | 156 | 4987 | - | - | - | - | - | - | - | - | 15 | 344 | - | - | - | - | - | - | 171 | 5331 |
| | PN-training-f | 7 | 269 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 5 | 162 | 12 | 431 |
| | Total | 179 | 5411 | - | - | - | - | - | - | - | - | 20 | 393 | 3 | 125 | 1 | 18 | 7 | 188 | **210** | **6135** |
| Training | PN-training-b | 148 | 1313 | - | - | - | - | - | - | - | - | 39 | 375 | - | - | - | - | - | - | 187 | 1688 |
| | PN-training-c | 6 | 340 | - | - | - | - | - | - | - | - | - | - | 9 | 772 | 7 | 530 | - | - | 22 | 1642 |
| | PN-training-d | 23 | 302 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 24 | 493 | 47 | 795 |
| | PN-training-e | 1419 | 46564 | - | - | - | - | - | - | - | - | 128 | 2631 | - | - | - | - | - | - | 1547 | 49195 |
| | PN-training-f | 71 | 2820 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 26 | 1009 | 97 | 3829 |
| | Total | 1667 | 51339 | - | - | - | - | - | - | - | - | 167 | 3006 | 9 | 772 | 7 | 530 | 50 | 1502 | **1900** | **57149** |

the training set. Details about the splitting are shown in Table 7.3.

The best way to prevent overfitting in neural networks is to train on more data. Therefore, we augment the training data by using various audio transformations from SoX [144], similar as in [145]. We consider the following two transformations to slightly modify the heart sound recordings:

- *Pitch*: Change the audio pitch without changing tempo.
- *Tempo*: Change the audio playback speed but not its pitch.

We modify the recordings from the training set with a pitch shift of ± *a semitone*, i.e. a fundamental frequency of 50 Hz varies with approximately ±3 Hz. We modify the time-scale of the recordings with ±10 %. In total, we get an augmented dataset consisting of 9500 recordings and 285745 beats, as shown in Table 7.4.

*Table 7.4: Augmented training set.*

| Effect | Parameters | # Recordings | # Beats |
|---|---|---|---|
| Clean | | 1900 | 57149 |
| Pitch | +semitone | 1900 | 57149 |
| Pitch | -semitone | 1900 | 57149 |
| Tempo | +10% | 1900 | 57149 |
| Tempo | -10% | 1900 | 57149 |
| Total | | 9500 | 285745 |

### 7.3.3 Labeling

Based on the hand-corrected annotations, we generated the labeling for the state sequence *S1 - systole - S2 - diastole*. Due to the shift of 20 ms in our frame-wise processing framework (see Section 7.4), we generated a label for each frame from the annotation information. In addition to the state labeling, we further added the label *no signal*, for areas with absent signal due to zero-padding. Figure 7.4 shows an example of a phonocardiogram (PCG) with the five labels.

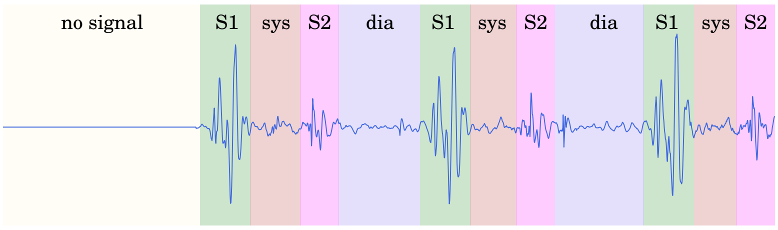


*Figure 7.4: Example of a PCG showing the five possible labels: no signal, first heart sound (S1), systole (sys), second heart sound (S2), and diastole (dia).*

## 7.4 Feature Extraction

We resampled the heart sound recordings to a sampling frequency of $f_s = 1\,\text{kHz}$ and removed direct current (DC) offset with a high-pass filter with a cut-off frequency of $f_c = 10\,\text{Hz}$ (relevant for PN-training e). We zero-padded the recordings according to the longest one in each set.

For the spectral features, we preprocessed all recordings with a short-time Fourier transform (STFT) using a Hamming window with window-size 80 ms ($\widehat{=}$ 80 samples) and 75 % overlap ($\widehat{=}$ frame-shifts of 20 ms or 20 samples). To exploit the spectral information of the heart sounds, we consider the following two types of features:

- *Spectrogram*: We extract 41-bin log magnitude spectrograms.
- *MFCCs*: MFCCs are used as features in various acoustic pattern recognition tasks, including heart sound classification [107] and heart sound segmentation [127].

  We extract 20 static MFCC coefficients, 20 delta MFCC coefficients ($\Delta$) and 20 acceleration MFCC coefficients ($\Delta^2$). This results in a 60-bin vector per frame. We use 20 mel bands within a frequency range of 0-500 Hz. The width used to calculate the delta and acceleration MFCC coefficients is 9 frames.

Furthermore, similar as for the LR-HSMM-method [134], we extract feature vectors for 20 ms-frames with the following features:

- *Envelope features* [134]: Homomorphic envelope, Hilbert envelope, wavelet envelope and power spectral density.

All features were normalised to zero-mean unit variance using the training corpus.

## 7.5 Evaluation Metrics

We perform an event-based evaluation of the results. We define an event as correctly detected if its temporal position overlaps with the one of an identically labeled event in the hand annotated ground truth. We allow a tolerance for the temporal onset and offset of ±40 ms ($\widehat{=}$ ±two frame-shifts of 20 ms), respectively. We determine for all heart sound recordings:

- True positives ($TP$): Events, where system output and ground truth have a temporal overlap;
- False positives ($FP$): The ground truth indicates no event that the system outputs;
- False negatives ($FN$): The ground truth indicates an event that is not recognised by the system;

We evaluate the performance of the segmentation algorithm using Precision (Equation 7.1), Sensitivity (Equation 7.2), and F-score (Equation 7.3). The F-score should be large. For a more detailed description of the metrics, we refer to [146].

$$P_+ = \frac{TP}{TP + FP} \tag{7.1}$$

$$Se = \frac{TP}{TP + FN} \tag{7.2}$$

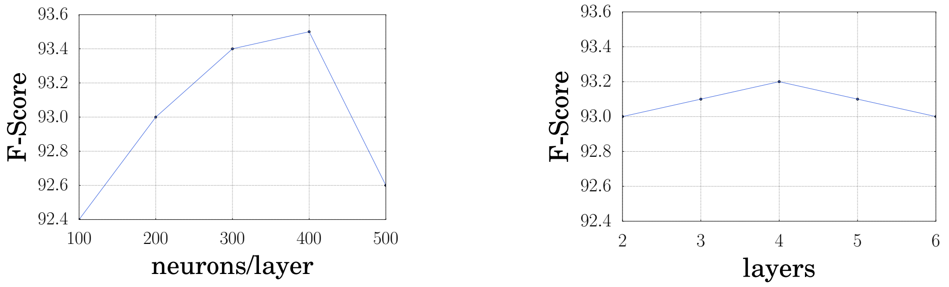
$$F_1 = 2 \cdot \frac{P_+ \cdot Se}{P_+ + Se} \tag{7.3}$$

## 7.6 Experiments and Results

For the experiments[5], we built a single multi-label classification system. We initialize the models with orthogonal weights [147] and use a softmax output gate as output layer. For optimizing the *cross-entropy error (CEE)* objective, we use Adam [148]. We perform early stopping, where we train each model for 200 epochs and use the parameter setting that causes the smallest validation error for the evaluation of the model. The reported scores are the average values over the events *S1*, *systole*, *S2*, and *diastole*. In addition to the average values, we report the scores for each event independently on the test set for the best setup. The reported scores are results of the validation set, except for the evaluation of the final setup on the test set in Section 7.6.6.

### 7.6.1 Comparison of GRNN Network Size

We initiate our experiments with finding an appropriate network size by using GRNNs and MFCC features. We use rectifier activations for the gated recurrent units. Figure 7.5 shows the results with varying number of neurons per hidden layer and varying number of hidden layers per model. For a 2-hidden layer GRNN, we achieved the best score of $F_1 = 93.5\,\%$ with 400 neurons/layer. For a GRNN with hidden layers of 200 neurons, we achieved the best score of $F_1 = 93.2\%$ with 4 hidden layers. Due to the small difference regarding the F-score, we choose a network size in favor of faster training. For the subsequent experiments, we fix the model size to 2 hidden layers, and 200 neurons per layer.



(a) Comparing network size

(b) Comparing network depth

*Figure 7.5: Comparison of network size: (a) shows the F-scores for a GRNN using* $\{2, ..., 6\}$ *hidden layer of 200 neurons. (b) shows the F-score for a GRNN with two hidden layer using* $\{100, 200, ..., 500\}$ *neurons per layer.*

[5] We conducted experiments using Python with Theano, and CUDA for GPU computing.

### 7.6.2 Comparison of GRNN Activation Functions

Table 7.5 shows the results for different activation functions. In particular, we use *sigmoid, tanh*, and *rectifier* non-linearities. Again, we use (2 hidden layer, 200 neurons per layer) GRNNs and MFCC features. Rectifier functions achieve the best average score, i.e. $F_1 = 93.0\,\%$. This is consistent with the literature [149].

*Table 7.5: Comparing different activation functions using GRNNs.*

| Model | Features | Activation | $P_+(\%)$ | $Se(\%)$ | $F_1(\%)$ |
|---|---|---|---|---|---|
| GRNN | MFCCs | sigmoid | 91.4 | 92.1 | 91.7 |
| GRNN | MFCCs | tanh | 92.4 | 93.3 | 92.8 |
| GRNN | MFCCs | rectifier | 92.2 | 93.8 | **93.0** |

### 7.6.3 Comparison of RNN Architectures

Table 7.6 shows the results for different RNN architectures. We compare different models, i.e. RNNs, LSTMs, GRNNs, and their bidirectional versions, using MFCC features. The model size for the conventional models is 2 hidden layers and 200 neurons per layer. For the bidirectional models, we use 2 hidden layers and 100 neurons for the forward and backward layers, respectively. The bidirectional long short-term memory (BiLSTM) slightly outperforms the other models, by achieving an average $F$-Score of $F_1 = 94.1\,\%$. Due to the small difference between the BiLSTM and the BiGRNN, we choose the less complex BiGRNN for the subsequent experiments.

*Table 7.6: Comparison of different recurrent neural networks architectures using MFCC features.*

| Model | Features | $P_+(\%)$ | $Se(\%)$ | $F_1(\%)$ |
|---|---|---|---|---|
| RNN | MFCCs | 90.2 | 93.1 | 91.6 |
| LSTM | MFCCs | 91.8 | 93.3 | 92.5 |
| GRNN | MFCCs | 92.2 | 93.8 | 93.0 |
| BiRNN | MFCCs | 91.8 | 94.5 | 93.1 |
| BiLSTM | MFCCs | 93.5 | 94.8 | 94.1 |
| BiGRNN | MFCCs | 92.8 | 94.5 | **93.7** |

### 7.6.4 Comparison of BiGRNN Input Features

Table 7.7 shows the results for BiGRNNs with MFCCs, spectrograms, envelope features, and their combinations. Best results are obtained with spectrograms, envelope features, and their combination. The envelope features already show promising results in combination with the LR-HSMMs-method. For this reason, and with the assumption that spectrograms render the segmentation more robust against artefacts, we use the combination of spectrogram and envelope features for the subsequent experiments.

*Table 7.7: Comparison of MFCCs, spectrogram, and envelope features.*

| Model | Features | $P_+$(%) | $Se$(%) | $F_1$(%) |
|---|---|---|---|---|
| BiGRNNs | MFCCs | 92.8 | 94.5 | 93.7 |
| BiGRNNs | Spectrogram | 95.0 | 95.7 | 95.4 |
| BiGRNNs | Envelope | 95.0 | 95.9 | 95.4 |
| BiGRNNs | MFCCs + Envelope | 93.7 | 94.6 | 94.2 |
| BiGRNNs | Spectrogram + Envelope | 94.9 | 95.8 | **95.4** |

## 7.6.5 Comparison of Different Regularizers

Table 7.8 shows the results using a BiGRNN with different regularization approaches. For dropout, we dropped units in the hidden layers during training with a probability of $p = 0.1$. This value achieved the best results among $p \in \{0.1, 0.5, 0.7, 0.9\}$. For VAT, we used the parameter setting of $\lambda = 0.1$, $\epsilon = 0.1$, and $I_p = 1$. For noise injection (cf. Section 6.2.3), we added zero mean Gaussian noise with standard deviation $\sigma = 0.025$ and magnitude $m = 0.25$. For data augmentation with audio transformations, we used the augmented training set for training (see Table 7.4). All regularization methods, except for data augmentation with audio transformations, improved the F-score. In particular, with dropout, we achieve the best result of $F_1 = 96.1\,\%$.

*Table 7.8: Comparison of different regularization methods.*

| Model | Regularizer | Parameters | $P_+$(%) | $Se$(%) | $F_1$(%) |
|---|---|---|---|---|---|
| BiGRNN | - | - | 94.9 | 95.8 | 95.4 |
| BiGRNN | VAT | $\lambda = 0.1$,<br>$\epsilon = 0.1$,<br>$\mathrm{I}_p = 1$ | 95.3 | 96.1 | 95.7 |
| BiGRNN | Dropout | p = 0.1 | 95.8 | 96.3 | **96.1** |
| BiGRNN | Noise Injection | $\sigma = 0.025$,<br>m = 0.25 | 95.4 | 95.8 | 95.7 |
| BiGRNN | Audio Transformations | - | 95.0 | 95.7 | 95.4 |

## 7.6.6 Evaluation of the Final Setup on the Test Set

Table 7.9 shows the results for the best setup (i.e. BiGRNN, 2 hidden layers, 200 neurons per layer, rectifier activations, spectrogram+envelope features, dropout regularization) evaluated on the test set. In addition to the metrics from the previous sections, we report in detail the numbers of reference states $N_{ref}$ (ground truth), system states $N_{sys}$ (BiGRNN-method), true positives $N_{TP}$, false negatives $N_{FN}$, and false positives $N_{FP}$ for each event, respectively.

*Table 7.9: Detailed results per event evaluated with the final setup on the test set.*

| Event | $N_{ref}$ | $N_{sys}$ | $N_{TP}$ | $N_{FN}$ | $N_{FP}$ | $P_{+}(\%)$ | $Se(\%)$ | $F_1(\%)$ |
|---|---|---|---|---|---|---|---|---|
| S1 | 21115 | 21271 | 20659 | 456 | 612 | 97.1 | 97.8 | 97.5 |
| Systole | 21200 | 21453 | 20267 | 933 | 1186 | 94.5 | 95.6 | 95.0 |
| S2 | 21073 | 21229 | 20102 | 971 | 1127 | 94.7 | 95.4 | 95.0 |
| Diastole | 21385 | 21758 | 20283 | 1102 | 1475 | 93.2 | 94.8 | 94.0 |
| Average | | | | | | 94.9 | 95.9 | 95.4 |

### 7.6.7 Comparison with the LR-HSMM-method

For this experiment, we remove recordings from the training and test set containing areas with no signal, because the LR-HSMM is limited to the detection of four events in the order of *S1 - systole - S2 - diastole*. This results in 1810 recordings for the training set and 744 recordings for the test set.

For the LR-HSMM, we preprocess the recordings with resampling to $f_s = 1\,\text{kHz}$ and high-pass filtering with a cut-off frequency of $f_c = 10\,\text{Hz}$ (cf. Section 7.4). We process the audio signals with frames of 20 ms. We train the LR-HSMM by using the four feature types provided: homomorphic envelogram, Hilbert envelope, wavelet envelope, and PSD [134].

Table 7.10 shows the results achieved with the BiGRNN (final setup) compared with the LR-HSMM method.

Figure 7.6 shows eight examples of automatically segmented heart sound recordings (snippets of four seconds each). In each subfigure, we show the hand annotated ground truth (GT), the segmentation with the LR-HSMM method and the segmentation with the BiGRNN. We show five recordings from *PN-Training-a* (Figure 7.6a to 7.6e), two recording from *PN-Training-b* (Figure 7.6f and 7.6g), and one recording from *PN-Training-e* (Figure 7.6h). For the visualization, we normalised each heart sound recording according to its maximum amplitude.

## 7.7 Discussion and Conclusion

In our experiments, we compare *vanilla* RNNs, LSTMs, GRNNs, and their bidirectional implementations, with BiGRNNs outperforming the rest. In subsequent experiments, we find the final setup using spectrogram and envelope features with a regularized BiGRNN. The network consists of 2 hidden layers with 200 neurons each and rectifier activations (except for the last layer). Regularization with *dropout* achieves the best result. Data augmentation with *audio transformations* does not result in any improvement.

In Section 7.6.7, we compare our proposed method with the state-of-the-art, the LR-HSMM-method. The BiGRNN-method performs on par with the LR-HSMM-method with an overall F-score of $F_1 = 95.6\,\%$ (cf. Table 7.10). We have to remark that this is not a completely fair comparison, because the ground truth annotations, although trained on less data (i.e. PN-training-a) and manually corrected, were generated with the LR-HSMM-method. This may introduce bias

*Table 7.10: Comparison of our BiGRNN with the LR-HSMM-method [134], evaluated on 744 recordings from the test set.*

| Challange set | Disease | #Recordings | #Beats | BiGRNN | | | LR-HSMM | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $P_+$(%) | $Se$(%) | $F_1$(%) | $P_+$(%) | $Se$(%) | $F_1$(%) |
| PN-training-a | Normal | 112 | 4270 | 95.6 | 96.7 | 96.1 | 96.4 | 96.0 | 96.2 |
| | MVP | 119 | 4402 | 93.1 | 93.8 | 93.4 | 91.2 | 91.6 | 91.4 |
| | Benign | 113 | 4163 | 96.4 | 97.2 | 96.8 | 96.8 | 96.8 | 96.8 |
| | AD | 13 | 425 | 86.6 | 93.2 | 89.7 | 95.5 | 95.5 | 95.5 |
| | MPC | 23 | 932 | 89.6 | 92.7 | 91.1 | 91.3 | 91.8 | 91.6 |
| | All | 380 | 14192 | 94.4 | 95.6 | **95.0** | 94.5 | 94.6 | **94.6** |
| PN-training-b | Normal | 132 | 1256 | 92.8 | 94.4 | 93.6 | 96.8 | 96.1 | 96.4 |
| | CAD | 29 | 238 | 79.9 | 82.8 | 81.3 | 94.6 | 93.5 | 94.0 |
| | All | 161 | 1494 | 90.6 | 92.5 | **91.6** | 96.5 | 95.7 | **96.1** |
| PN-training-e | Normal | 201 | 4981 | 98.6 | 98.6 | 98.6 | 98.5 | 98.3 | 98.4 |
| | CAD | 2 | 19 | 68.1 | 69.4 | 68.7 | 85.9 | 85.9 | 85.9 |
| | All | 203 | 5000 | 98.5 | 98.5 | **98.5** | 98.4 | 98.2 | **98.3** |
| All | | 744 | 20686 | 95.1 | 96.1 | **95.6** | 95.6 | 95.5 | **95.6** |

towards the LR-HSMM-method. Furthermore, the hand annotated ground truth is not always correct (cf. Figure 7.6a and 7.6h), also being in favor of the LR-HSMM-method and in general causing distortion in the scores.

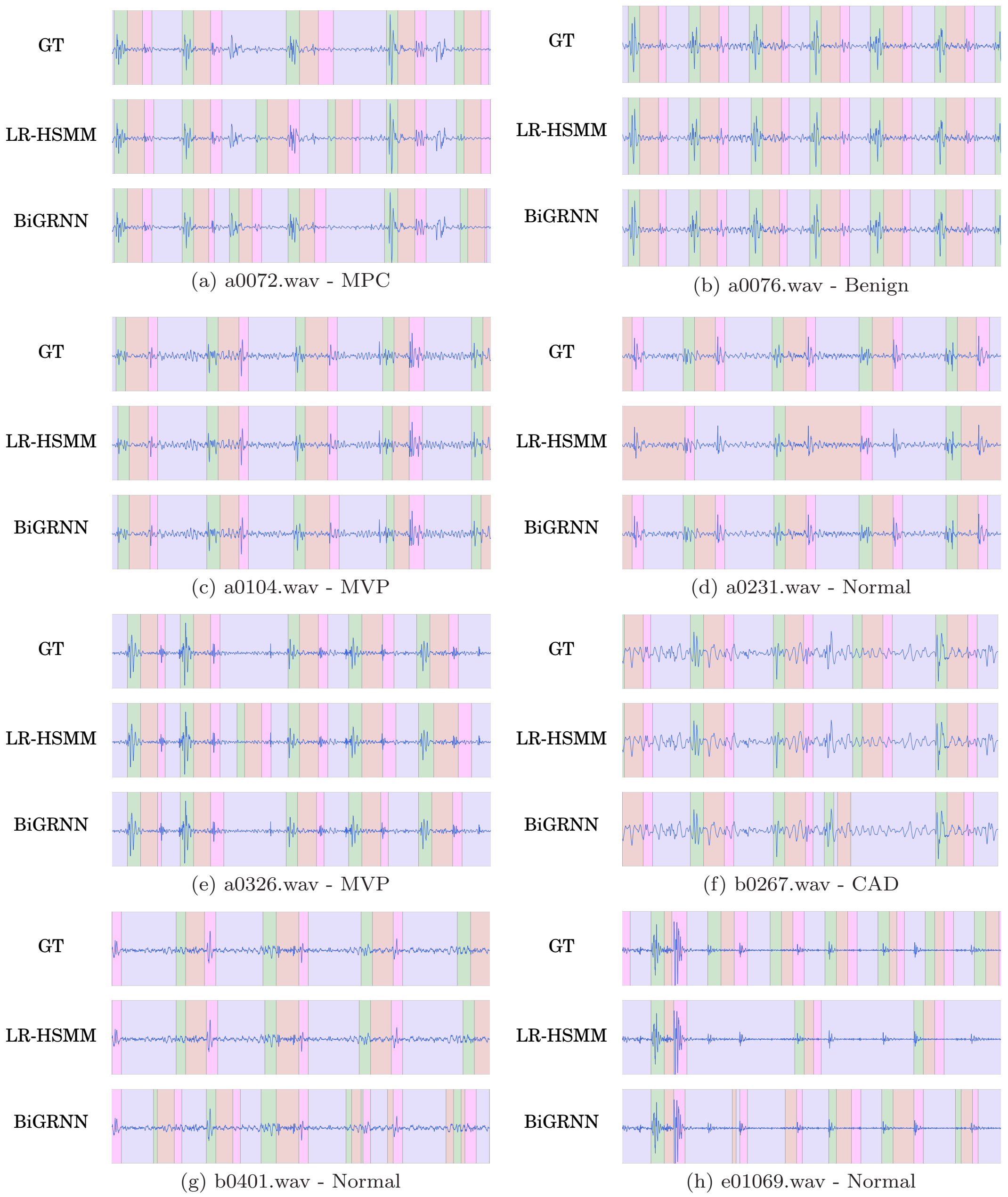
Table 7.10 shows detailed results for the test data in terms of PN-training sets and diseases. We observe that the BiGRNN-method outperforms the LR-HSMM-method for PN-training-a and PN-training-e, but performs worse for PN-training-b. Regarding the diseases in PN-training-a, only for MVP the BiGRNN-method outperforms the LR-HSMM-method, and for benign murmurs (Benign) both methods perform on par. For PN-training-b, the LR-HSMM-method outperforms the BiGRNN-method for normal and CAD recordings. For normal recordings of PN-training-e, the BiGRNN-method outperforms the LR-HSMM-method. The LR-HSMM-method is distinctly better than the BiGRNN-method for the two recordings of CAD in PN-training-e.

The 2016 PhysioNet/CinC Challenge data does not provide any labeling for cardiac arrhythmia. According to [150], mitral valve prolapse is a source of arrhythmias. We refer to the results reported for MVP, with the BiGRNN-method ($F_1 = 93.4\,\%$) outperforming the LR-HSMM-method ($F_1 = 91.4\,\%$). Moreover, we visually inspected all test recordings of MVP and found 20 recordings with cardiac arrhythmia. On this selected set of recordings the BiGRNN-method ($F_1 = 87.2\,\%$) outperforms the LR-HSMM-method ($F_1 = 75.7\,\%$). An example for cardiac arrhythmia for MVP is shown in Figure 7.6e.

The examples in Figure 7.6 illustrate some observations for both segmentation methods, and also for the ground truth annotations. Figure 7.6b and 7.6c show examples, where both methods perform well. In Figure 7.6d, we observe that the LR-HSMM-method skips every second *S2*, and detects every second *S1* as *S2*. Figure 7.6g shows some segmentation errors for the BiGRNN method. Figure 7.6a, 7.6h and 7.6f are examples, where the ground truth labeling is partially incorrect. In Figure 7.6h, we further observe that both methods achieve partially incorrect segmentation results. Figure 7.6e shows an example for the failure of the LR-HSMM method for irregularity of the temporal occurrence of the events.

In our experiments, the proposed BiGRNN-method achieves performance on par with the LR-HSMM-method. We successfully show state-of-the-art performance without directly incorporating a priori information of the state durations. The proposed method is easily extendable to the detection of extra heart sounds (third and fourth heart sound), heart murmurs, as well as other acoustic events. However, this would require appropriate training data, i.e. heart sound recordings containing the additional events and their proper labeling. In a practical sense, our method features further advantages. Without preprocessing, it can easily handle absence of the signal, noise, and irregularity of the temporal occurrence of the events (like in cardiac arrhythmia).

The proposed method represents a general solution for the detection of different kinds of events in heart sound recordings. The method is easily extendable to the detection of extra heart sounds (third and fourth heart sound), heart murmurs, as well as other acoustic events. This, however, requires appropriate training data with *thorough* labeling of the events and further experiments.



(a) a0072.wav - MPC

(b) a0076.wav - Benign

(c) a0104.wav - MVP

(d) a0231.wav - Normal

(e) a0326.wav - MVP

(f) b0267.wav - CAD

(g) b0401.wav - Normal

(h) e01069.wav - Normal

*Figure 7.6: Legend:* 🟩 *S1;* 🟫 *systole;* 🟪 *S2;* 🟦 *diastole.*
*Examples of automatically segmented heart sound recordings (snippets of four seconds each) from the test set. In each subfigure, the first plot corresponds to the hand annotated GT, the second to the LR-HSMM method and the third to the proposed method (BiGRNN).*

# 8 Event Detection in Lung Sounds

In this chapter, we apply the framework from Chapter 7 to lung sounds. In particular, we use it for the detection of crackles and breathing phase events in single-channel lung sound recordings.

We organised this chapter as follows. A short introduction is given in Section 8.1. Adopted from Section 7.2, we present the proposed event detection method in Section 8.2. We describe the data acquisition and the recording material in Section 8.3. Experimental results are shown in Section 8.6. We discuss the results and conclude this chapter in Section 8.7.

*This chapter was published in the proceedings of the 40th Annual Conference of the IEEE Engineering in Medicine and Biology Society 2018 under the title of "Crackle and Breathing Phase Detection in Lung Sounds with Deep Bidirectional Gated Recurrent Neural Networks" [151]. As minor modifications, we changed some wordings and added a new figure.*

## 8.1 Introduction

For adventitious sound detection, the temporal position of the event within the breathing phase is of interest. Therefore, besides the detection of the adventitious sound itself, breathing phase detection (BPD) is also needed. In lung sound research, BPD is usually accomplished with information gained by simultaneous airflow measurement [152]. This, however, is inconvenient for long term monitoring [153] or not feasible due to hardware limitations (e.g. electronic stethoscopes). Existing approaches for acoustic BPD are limited to defined recording locations, by either using tracheal sounds [153] or a combination of tracheal sounds and lung sounds [154,155]. Approaches for crackle detection are threshold-based classifiers [156, 157] or machine learning methods, using algorithms like SVMs, k-NN, and MLP [158]. Based on the current state-of-the-art, we face several challenges: (i) We seek for a unified solution for both crackle and breathing phase detection. (ii) We aim to be robust against disturbing sound sources. (iii) We seek for robustness in terms of recording position, since lung sounds characteristics vary with the recording location over the chest.

To this end, we introduce an event detection approach with GRNNs [96, 97] for crackle and breathing phase detection in single-channel lung sound recordings. In particular, for the first time, we propose a multi-label classification system with BiGRNNs, using spectral features. With the proposed method, we exploit spectral information and temporal dependencies of the lung sounds. We use lung sound recordings from lung-healthy subjects and patients with IPF (cf. Section 5.3).

Due to the limitations from our data, we focus on crackle detection, although our proposed method is generally applicable to the detection of all kinds of adventitious sounds. In experiments, we report event based metrics and visualise examples of automatically detected breathing phases and crackles in lung sound recordings.

## 8.2 Audio Processing Framework

The essential steps of the event detection framework are shown in Figure 8.1. Given a raw single-channel lung sound recording $\mathbf{x}_t = [x_1, \ldots, x_T]$ in the time domain, we extract a sequence of feature frames $\mathbf{x}_f \in \mathbb{R}^D$, where $D$ is the dimension of the feature vector and $f \in \{1, ..., F\}$ is the frame index, with $F$ indicating the number of frames.

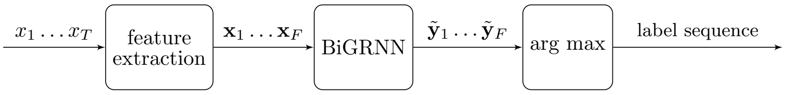

$x_1 \ldots x_T$
feature extraction
$\mathbf{x}_1 \ldots \mathbf{x}_F$
BiGRNN
$\tilde{\mathbf{y}}_1 \ldots \tilde{\mathbf{y}}_F$
arg max
label sequence

*Figure 8.1: Event detection framework.*

These feature frames are processed by a multi-label BiGRNN with a softmax output layer. The index of the maximum value of the real-valued output vector $\tilde{\mathbf{y}}_f$ determines the event class per frame, resulting in a sequence of frame labels as output. Consecutive identical frame labels are grouped as one event. A schematic illustration of the frame-wise single-channel lung sound processing framework is shown in Figure 8.2.

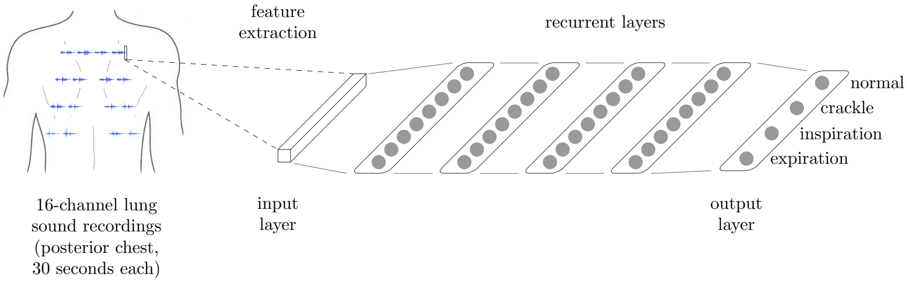


*Figure 8.2: Frame-wise single-channel lung sound processing framework with a recurrent neural network.*

## 8.3 Material - Lung Sound Recordings

### 8.3.1 Lung Sound Recordings

We used a subset[6] of the recordings from our multi-channel lung sound database from Section 5.3. Table 8.1 shows the resulting list of subjects. It contains ten lung-healthy subjects and five with IPF in an advanced stage. For each subject, we used 16-channel lung sound recordings at two different airflow rates, with 3-8 breathing cycles within 30 seconds, respectively. This results in 32 single-channel recordings of 30 seconds per subject and 480 single-channel recordings in total, at varying airflow rates and recording positions.

*Table 8.1: Subjects list for the event detection dataset, including given maximum inspiratory airflow values during the measurements.*

| Subject # | Gender | Age | Height | Weight | BMI | Category | Max. Insp. Airflow [l/s] |
|---|---|---|---|---|---|---|---|
| 1 | male | 27 | 178 | 78 | 24.6 | lung-healthy | 1.0 & 1.5 |
| 2 | male | 42 | 167 | 62 | 22.2 | lung-healthy | 1.0 & 1.3 |
| 3 | male | 26 | 189 | 75 | 21.0 | lung-healthy | 1.0 & 1.2 |
| 4 | male | 30 | 193 | 74 | 19.9 | lung-healthy | 1.2 & 1.5 |
| 5 | male | 27 | 173 | 85 | 28.4 | lung-healthy | 1.0 & 1.3 |
| 6 | male | 23 | 193 | 70 | 18.8 | lung-healthy | 0.6 & 1.0 |
| 7 | male | 41 | 180 | 97 | 29.9 | lung-healthy | 0.5 & 1.2 |
| 8 | male | 28 | 172 | 82 | 27.7 | lung-healthy | 0.5 & 1.0 |
| 9 | male | 53 | 180 | 80 | 24.7 | lung-healthy | 0.7 & 1.7 |
| 12 | female | 24 | 172 | 73 | 24.7 | lung-healthy | 0.7 & 1.3 |
| 17 | male | 76 | 184 | 92 | 27.2 | IPF | 0.8 & 1.0 |
| 18 | male | 60 | 175 | 82 | 26.8 | IPF | 1.0 & 2.0 |
| 19 | male | 79 | 175 | 75 | 24.5 | IPF | 1.0 & 1.2 |
| 20 | male | 74 | 187 | 83 | 23.7 | IPF | 1.0 & 1.2 |
| 23 | female | 76 | 158 | 53 | 21.2 | IPF | 0.5 & 1.0 |

### 8.3.2 Annotation of Acoustic Events

In all lung sound recordings, we annotated the temporal onset and offset positions of the events *inspiration*, *expiration*, and *crackles*. We labeled *crackles* by grouping consecutive crackles as one event and manually annotating their temporal position, i.e. this was not just done for a whole multi-channel recording once, but for each of the 16 recordings individually. We generated the labels for the breathing phases using the airflow signal. Firstly, we smoothed the airflow signal by lowpass filtering (cut-off frequency $f_c = 3\,\text{Hz}$). The zero-crossing positions of the airflow signal provided the onset positions of the breathing phases. The sign of the signal values indicated the breathing phase label. Table 8.2 gives an overview on the number of subjects, recordings and events in the dataset.

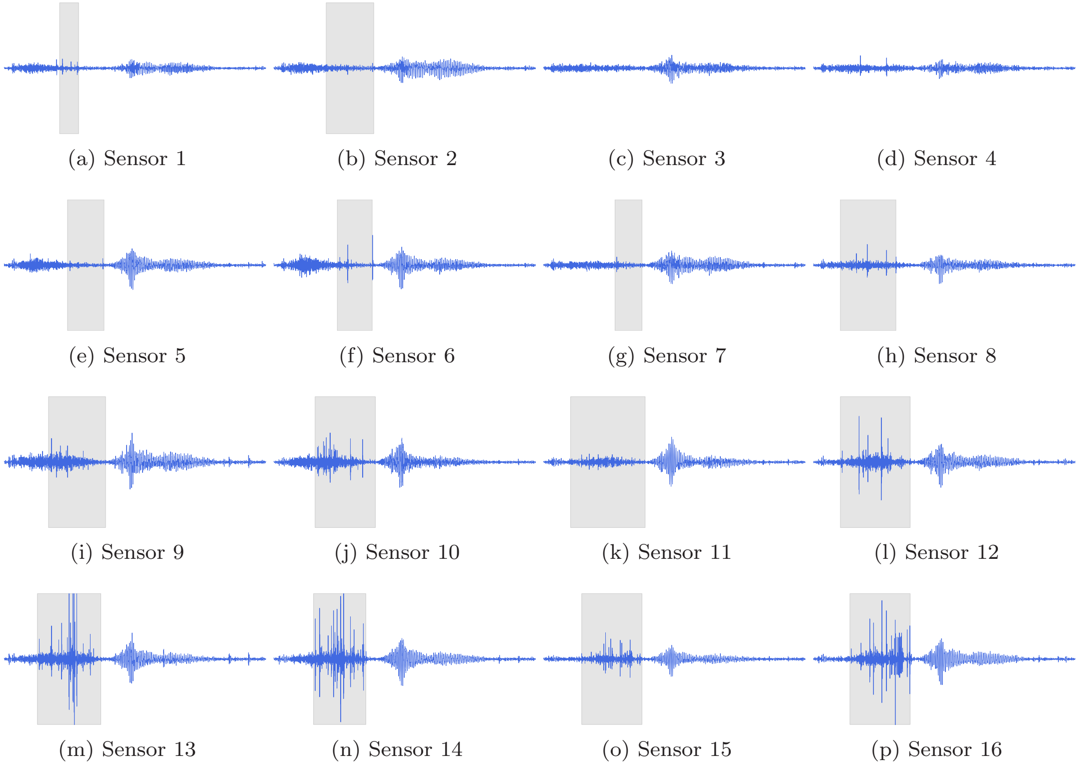
Characteristic adventitious sounds for IPF are mid to late inspiratory crackles [38,86]. Figure 8.3 shows one example of a 16-channel lung sound recording related to IPF. The sensors are numbered

[6] At the time of writing of the underlying conference paper, this was the amount of data available.

*Table 8.2: Number of subjects, recordings and events in the dataset.*

| # Subjects | | # Recordings | # Events | | |
|---|---|---|---|---|---|
| Healthy | IPF | | Inspiration | Expiration | Crackles |
| 10 | 5 | 480 | 4656 | 4720 | 1339 |

according to the pattern on the recording device (cf. Figure 3.2). In each recording, the temporal onset and offset positions of inspiratory crackles (grouped consecutive crackles) are marked. Strong crackles can be observed in basal areas (bottom rows) and weaker or no crackles in apical areas (top rows).


(a) Sensor 1 (b) Sensor 2 (c) Sensor 3 (d) Sensor 4

(e) Sensor 5 (f) Sensor 6 (g) Sensor 7 (h) Sensor 8

(i) Sensor 9 (j) Sensor 10 (k) Sensor 11 (l) Sensor 12

(m) Sensor 13 (n) Sensor 14 (o) Sensor 15 (p) Sensor 16

*Figure 8.3: Example of a 16-channel lung sound recording (one full breathing cycles) from a subject with IPF. In each subfigure, the temporal positions of the manually annotated consecutive crackles are marked.*

## 8.4 Feature Extraction

We process the lung sound recordings with a sampling frequency of $f_s = 16\,\text{kHz}$. All recordings are processed with a STFT using a Hamming window using a window-size of 32 ms ($\widehat{=}$ 512 samples) and 12 ms overlap ($\widehat{=}$ frame-shifts of 20 ms). To exploit the spectral information of the lung sounds, similar as in [31], we extract the following types of features:

- *MFCCs*: We extract 20 static coefficients, 20 delta coefficients ($\Delta$), and 20 acceleration coefficients ($\Delta^2$). We use 40 mel bands within a frequency range of 0-8000 Hz. The width used to calculate the delta and acceleration coefficients is 9 frames. This results in a 60-bin vector $\mathbf{x}_f$ per frame.
- *Spectrogram*: We extract 257-bin log magnitude spectrograms.

## 8.5 Evaluation Metrics

We perform an event-based evaluation of the results [146]. An event is defined as correctly detected, if its temporal position overlaps with the one of an identically labeled event in the ground truth. We allow a tolerance for the temporal onset and offset of ±0.5 s, respectively. For all lung sound recordings, we determine:

- True positives ($TP$): Events, where system output and ground truth have a temporal overlap;
- False positives ($FP$): The ground truth indicates no event but the system recognises an event;
- False negatives ($FN$): The ground truth indicates an event that is not recognised by the system.

We use Precision (Equation 7.1), Sensitivity (Equation 7.2), and F-score (Equation 7.3) to evaluate the performance of the event detection algorithm.

## 8.6 Experiments and Discussion

We build a single multi-label classification system for our experiments[7]. We use a BiGRNN (cf. Section 6.1.2) consisting of two hidden layers with 100 neurons for the forward and backward layers, respectively. The output layer is split into two softmax layers, one for the outputs *normal* and *crackle*, the other one for *inspiration* and *expiration*. The activation functions in the hidden layers are rectifier non-linearities. We initialise the models with orthogonal weights [147]. For optimizing the *CEE* objective, we use Adam [148]. We use dropout [105] for regularization, applied to the hidden layers with a dropout probability of $p = 0.5$.

*Table 8.3: Dataset splitting in terms of subjects and diseases for one fold of 5-fold cross-validation.*

| Dataset | # Subjects | | # Recordings |
|---|---|---|---|
| | Healthy | IPF | |
| Training | 7 | 3 | 320 |
| Validation | 1 | 1 | 64 |
| Test | 2 | 1 | 96 |
| Total | 10 | 5 | 480 |

[7] We conducted experiments using Python with Theano, and Compute Unified Device Architecture (CUDA) for graphics processing unit (GPU) computing.

Due to few data samples, we use 5-fold cross-validation, with the splitting of the dataset shown in Table 8.3. We assign each subject exclusively to either training, validation, or test set. For each fold, we apply early stopping, i.e. we train each model for 100 epochs and use the parameter setting that causes the smallest validation error to process the test set data.

Table 8.4 shows the results for MFCCs and spectrogram features for each event. We report the micro-average of the scores from the five folds, evaluated on the test set. For all events MFCCs outperform spectrogram features in terms of the F-score.

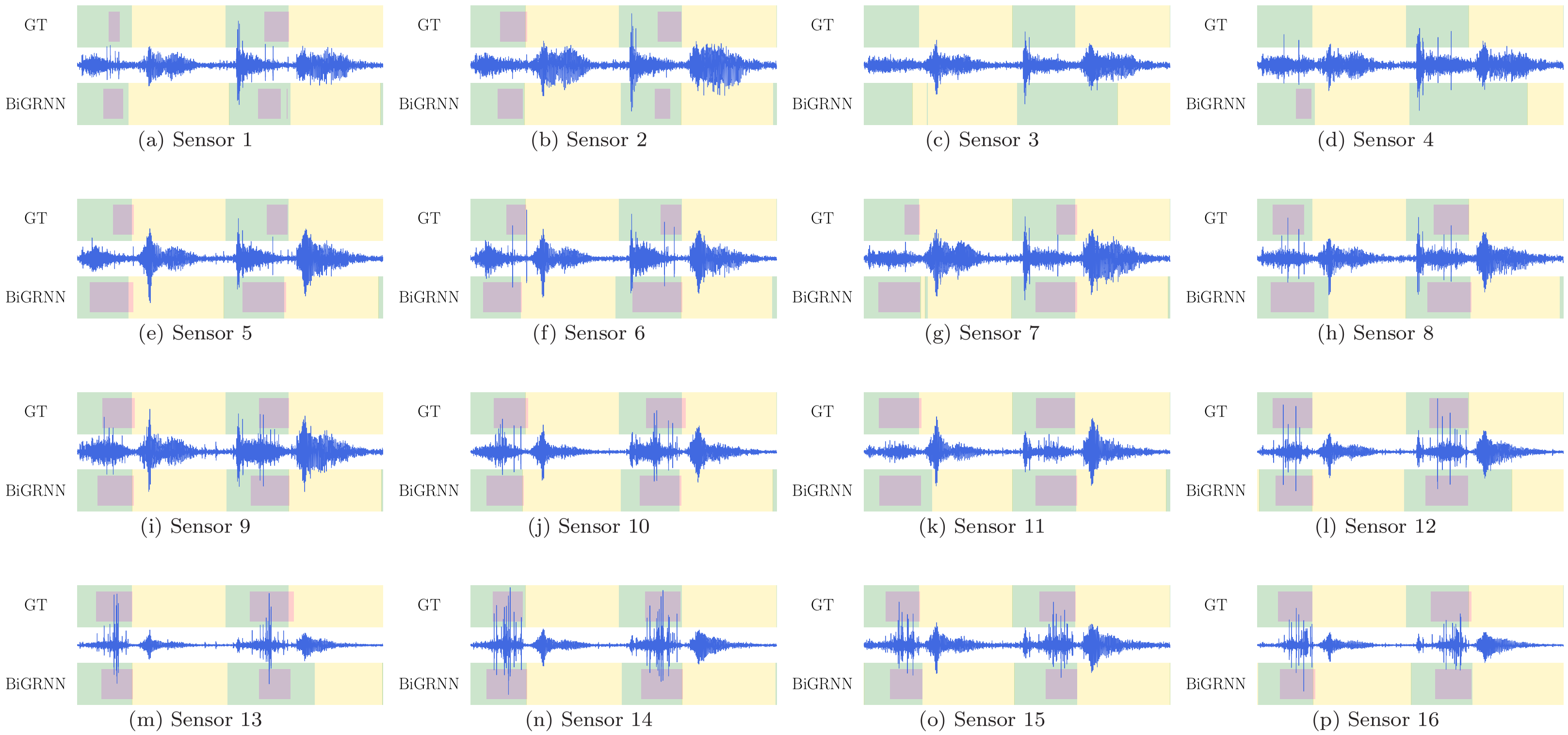
*Table 8.4: Results per event for MFCCs and spectrogram features.*

| Event | Feature | $P_+$(%) | $Se$(%) | $F_1$(%) |
|---|---|---|---|---|
| Inspiration | MFCCs | 83.7 | 90.5 | **87.0** |
| | Spectrogram | 86.6 | 86.7 | 86.7 |
| Expiration | MFCCs | 81.4 | 88.1 | **84.6** |
| | Spectrogram | 82.9 | 83.9 | 83.4 |
| Crackles | MFCCs | 72.6 | 71.5 | **72.1** |
| | Spectrogram | 71.6 | 65.6 | 68.5 |

Figure 8.4 shows examples of automatically detected events in a 16-channel lung sound recording (two breathing cycles) from a subject with IPF. The sensor positions correspond with the pattern from the recording device (see Figure 3.2). In each subfigure, we show the GT and the detected events with the BiGRNN. We normalised each of the depicted lung sound recordings according to its maximum amplitude. We observe loud crackles in the basal area (bottom rows) and weaker or no crackles in the apical area (top rows).

## 8.7 Discussion and Conclusion

In general, we observe for the event detection robustness regarding the contamination of the lung sound recordings with noise, bowel, and heart sounds. In some recordings with shallow breathing, no or weak lung sounds are audible in the lowest sensor row (Sensors 13 to 16), which results in failure of our method, especially affecting the performance of breathing phase detection. For the automatic detection of crackles, we observe sensitivity to the manual labeling accuracy, i.e. some crackles are detected although not distinctly identified in the manual labeling (cf. Figure 8.4 - Sensor 4). Furthermore, some crackle events are automatically detected in lung-healthy subjects, which are present, but not manually annotated, because not related to IPF. Despite the low number of data samples used in our experiments, the results are encouraging.



(a) Sensor 1 (b) Sensor 2 (c) Sensor 3 (d) Sensor 4

(e) Sensor 5 (f) Sensor 6 (g) Sensor 7 (h) Sensor 8

(i) Sensor 9 (j) Sensor 10 (k) Sensor 11 (l) Sensor 12

(m) Sensor 13 (n) Sensor 14 (o) Sensor 15 (p) Sensor 16

*Figure 8.4: Legend:* 🟩 Inspiration; 🟨 Expiration; 🟥 Crackles.
*Examples of automatically detected events in a 16-channel lung sound recording (two breathing cycles) from a subject with IPF. In each subfigure, the GT labeling and the automatic labeling with our BiGRNN are illustrated.*

# 9 Multi-channel Lung Sound Classification

In this chapter, we present an approach for multi-channel lung sound classification, exploiting spectral, temporal, and spatial information. In particular, we propose a frame-wise classification framework to process full breathing cycles of multi-channel lung sound recordings with a convolutional recurrent neural network.

This chapter is structured as follows. After a short introduction in Section 9.1, we present our proposed multi-channel classification framework in Section 9.2. The evaluation metrics used in our experiments are described in Section 9.4. The experimental setup and the results are presented in Section 9.5. Finally, we discuss our findings and conclude this chapter in Section 9.6.

*This chapter is submitted for publication in the IEEE Journal of Biomedical and Health Informatics under the title of "Multi-channel Lung Sound Classification with Convolutional Recurrent Neural Networks" [90]. As minor modifications, we changed some wordings and added a new figure.*

## 9.1 Introduction

Several approaches to multi-channel lung sound classification exist. Because of the lack of a publicly available multi-channel lung sound database, research groups record lung sounds independently with different recording setups, i.e. differing in design, and number and position of sensors. A first approach to multi-channel lung sound analysis was the STG16 [159]. It enables 14-channel lung sound recording on the posterior chest, with two additional channels for the locations trachea and heart. Algorithms enable the detection and localization of different adventitious sounds. Another multi-channel recording device with 14-channel lung sound recording on the posterior chest, however with a different sensor arrangement than the STG16 [159], is presented in [12]. The authors of [12] explore a useful methodology for the classification of the three-class structure (healthy-obstructive-restrictive) in [160]. They model 14-channel pulmonary sound data using a second order vector autoregressive (VAR) model, and feed the estimated model parameter to SVM and GMM classifiers. A 25-channel lung sound recording device is used in [13], with a $5 \times 5$ sensor array attached on the posterior chest. The authors assess different parameterization techniques for multi-channel lung sounds for two-class classification (normal versus abnormal), such as PSD, the eigenvalues of the covariance matrix, the univariate autoregressive (UAR), and the multivariate autoregressive (MAR). Those methods are applied to construct feature vectors used as input to a supervised multilayer neural network.

In this chapter, we focus on the classification of isolated (multi-channel) lung sound recordings with one full breathing cycle each, where we exploit spectral, temporal, and spatial information of lung sounds. The fixed pattern for the LST arrangement of our recording front-end results in varying recording positions depending on the subject's physique. Therefore, we present a multi-channel lung sound classification framework, which renders exact recording positions dispensable. Inspired from computer vision, we propose a classification approach with CNNs [100], which we combine with RNNs [30, 137], resulting in a convolutional recurrent neural network (CRNN) [161, 162]. As already described in Section 7.1, RNNs are suitable architectures to process sequential input of variable length and learn temporal dependencies within the data. Another powerful neural network architecture are CNNs [100]. They are widely applied to audio classification tasks, including lung sound classification [33, 34]. Convolutional neural networks can be used as feature extractors by directly applying them to raw audio waveforms [145, 163]. Another approach is their usage after feature extraction, i.e. by processing spectrograms [161]. Our main contributions and results can be summarised as follows:

- We introduce a suitable architecture with CRNNs to multi-channel lung sound classification.
- We present experimental results, where we compare different neural network architectures for classification.

## 9.2 Audio Processing Framework

The proposed classification framework processes multi-channel lung sound recordings of one breathing cycle each.

### 9.2.1 Basic Processing Framework

The essential steps of our lung sound classification framework are shown in Figure 9.1. Given a raw single-channel lung sound recording $\mathbf{x}_t = [x_1, \ldots, x_T]$ in time domain, we extract a sequence of feature frames $\mathbf{x}_f \in \mathbb{R}^D$, where $D$ is the dimension of the feature vector. For multi-channel processing, we stack the feature vectors of the single channels to one feature vector.

These feature frames are processed by a multi-label (i.e. three classes: *healthy*, *pathological*, *no signal*) DNN with a softmax output layer. The real-valued output vectors $\tilde{\mathbf{y}}_f$ for all frames $F$ are summed up to $\tilde{\mathbf{y}}_{sum}$. The maximum value in $\tilde{\mathbf{y}}_{sum}$ determines the final class, where the class *no signal* is ignored.

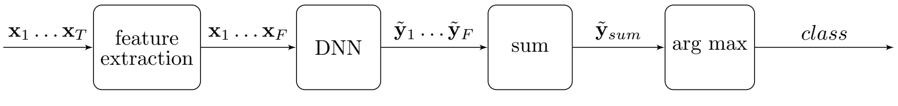

$\mathbf{x}_1 \ldots \mathbf{x}_T$
feature extraction
$\mathbf{x}_1 \ldots \mathbf{x}_F$
DNN
$\tilde{\mathbf{y}}_1 \ldots \tilde{\mathbf{y}}_F$
sum
$\tilde{\mathbf{y}}_{sum}$
arg max
*class*

*Figure 9.1: Multi-channel classification framework.*

For the feature extraction, we process the lung sound recordings with a sampling frequency of $f_s = 16\,\text{kHz}$. Due to convenience for the processing with the DNN, we zero-padded the recordings according to the longest one in each set. All recordings are processed with a STFT using a Hamming window with a window-size of 32 ms ($\widehat{=}$ 512 samples) and 12 ms overlap ($\widehat{=}$ frame-shifts of 20 ms). To exploit spectral information of the lung sounds, we extract 257-bin log magnitude spectrograms [91, 151]. The stacking of the individual feature vectors results in a 4112-dimensional feature vector. A schematic illustration of the basic frame-wise multi-channel lung sound processing framework with a multilayer perceptron or a recurrent neural network is shown in Figure 9.2.

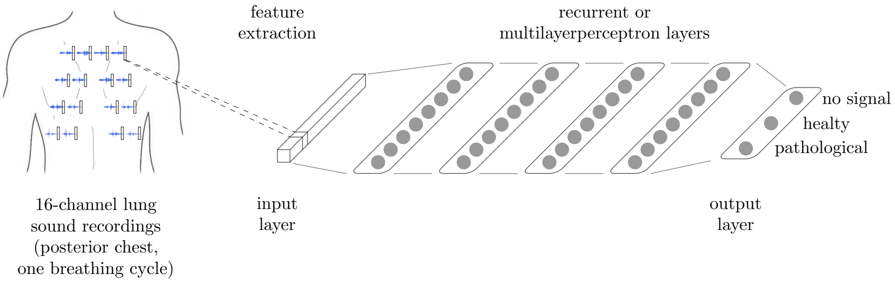


*Figure 9.2: Basic frame-wise multi-channel lung sound processing framework with a multilayer perceptron or a recurrent neural network.*

### 9.2.2 Extension with Convolutional Front-end

Our recording front-end features a fixed pattern for the LST arrangement (cf. Figure 3.2), resulting in varying recording positions depending on the subject's physique. Inspired from image processing, we propose an approach with convolutional neural networks to render exact recording positions dispensable.

Figure 9.3 shows the multi-channel lung sound processing framework with a recurrent neural network and a convolutional front-end, i.e. a CRNN. Compared to the basic processing framework (cf. Section 9.2.1), where the feature vectors of the 16-channel lung sounds are simply stacked to achieve one feature vector as input, we take the two-dimensional arrangement of the sensors into account. In each step, the feature frames of the 16 channels are combined in a 4x4 grid according to the sensor arrangement (cf. Figure 3.2), with the depth of the input 'image' corresponding to the dimension of the feature vector $D = 257$. At each time step, the input image is processed with convolutional layers and a subsampling layer, followed by (fully connected) recurrent layers, and the output layer. For the (fully connected) recurrent layers, we use a BiGRNN. Therefore, we call this architecture a convolutional bidirectional gated recurrent neural network (ConvBiGRNN). The first convolutional layer performs a dimensionality reduction with an $1 \times 1$ kernel [164]. In the second convolutional layer, a $3 \times 3$ kernel with stride and padding of one is used. After that, subsampling could be applied to reduce the information to four regions on the posterior chest (upper left, upper right, lower left, lower right). We determine the size of the network, the number of feature maps, and the subsampling experimentally in Section 9.5.

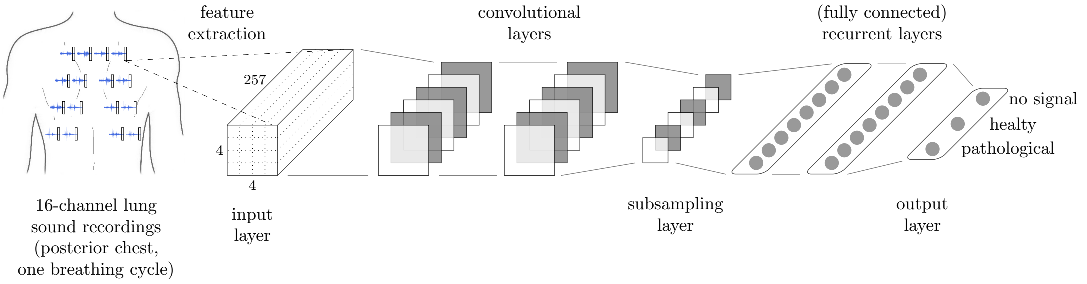


*Figure 9.3: Frame-wise multi-channel lung sound processing framework with a recurrent neural network and a convolutional neural network front-end.*

## 9.3 Material - Lung Sound Recordings

### 9.3.1 Lung Sound Recordings

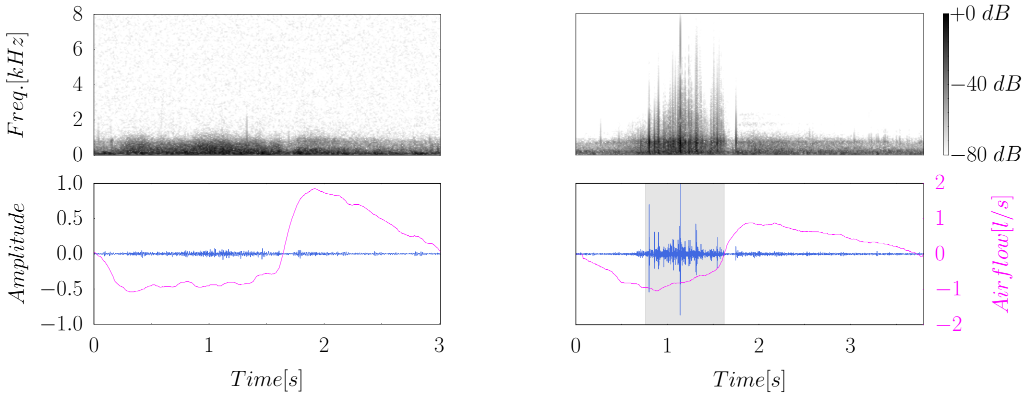
We used all of the recordings from our multi-channel lung sound database from Section 5.3. The dataset is summarised in Table 5.1. For each subject, we include 16-channel lung sound recordings at two different airflow rates. This results in two 16-channel lung sound recordings at varying airflow rates for each of the 23 subjects, with several breathing cycles within 30 seconds, respectively.

From all recordings, we extracted full breathing cycles by means of the airflow signal (cf. Figure 9.4). We smoothed the airflow signal with a low-pass filter (cut-off frequency $f_c = 3$ Hz). With the zero-crossing positions of the airflow signal, we determine the onset positions of the breathing phase. The sign of the signal values enables the distinction between inspiration and expiration. Table 9.1 gives an overview on the number of breathing cycles, and thus on the number of recordings (of one full breathing cycle) used in our experiments.

*Table 9.1: Number of subjects and recordings of one full breathing cycle.*

| # Subjects | | # Breathing Cycles | | |
|---|---|---|---|---|
| Healthy | IPF | Healthy | IPF | Total |
| 16 | 7 | 252 | 135 | 387 |

Figure 9.4 shows two examples of lung sound recordings, one from a lung-healthy subject (a) and one from a patient diagnosed with IPF (b). Depicted are one full breathing cycle of the audio waveform, the simultaneous airflow recording (lower plot), and the corresponding spectogram (upper plot), respectively. Both recordings are from the basal left lung area, i.e. from sensor 16 (cf. Figure 3.2). In (b), the inspiratory crackles related to IPF are marked. In the corresponding spectrogram, the intensity in the higher frequency range (above ≈ 2 kHz) is notable, when compared to (a).



(a) Lung-healthy

(b) IPF

*Figure 9.4: Two examples of lung sound recordings (one full breathing cycle each), one from a lung-healthy subject (a) and one from a patient diagnosed with IPF (b). In (b), the temporal position of consecutive crackles are marked with a gray box.*

## 9.4 Evaluation Metrics

We perform a breathing cycle-wise evaluation of the results. For all lung sound recordings, we determine:

- True positives ($TP$): correctly classified as pathological;
- False positives ($FP$): falsely classified as pathological;
- False negatives ($FN$): falsely classified as healthy.

We evaluate the performance of the classification algorithms using Precision (Equation 7.1), Sensitivity (Equation 7.2), and F-score (Equation 7.3). Precision $P_+$ provides information about how many of the recordings labeled as pathological are actually true. Sensitivity (or Recall) $Se$ provides information about how many of the pathological recordings are actually labeled as such. Due to the uneven class distribution, we use the F-score $F_1$ as an overall performance measure.

## 9.5 Experimental Setup and Results

We compare two different neural network architectures used within the *basic processing framework* from Section 9.2.1. As a baseline system, we use an MLP. As a second model, we choose a BiGRNN, similar as in Chapter 7 and Chapter 8. Furthermore, we compare the results with the ConvBiGRNN from Section 9.2.2.

For all three networks, we determine the optimal network size with grid search, i.e. we select the architecture resulting in the highest F-score on the validation set[8].

[8] We conducted experiments using Python with Theano, and CUDA for GPU computing.

The activation functions $g(\cdot)$ in the convolutional layer and in the hidden layers of the MLP and the BiGRNN are rectifier non-linearities. We initialize the models with orthogonal weights [147]. For regularization, we add a dropout layer [91, 105] with a dropout probably of $p = 0.5$ after every hidden layer of the MLP and after every hidden recurrent layer of the BiGRNN and the ConvBiGRNN. We optimize the *CEE* objective with Adam [148].

Due to few data samples, we use 7-fold cross-validation, with the recordings of each IPF subject appearing once in the test set. The splitting of the dataset is shown in Table 9.2. Each subject is assigned to either training, validation, or test set. We train the models for 400 epochs and apply early stopping, i.e. we use the parameter setting that causes the smallest validation error to process the test set data.

*Table 9.2: Dataset split in terms of subjects and diseases for one fold of 7-fold cross-validation.*

| Dataset | # Subjects | |
|---|---|---|
| | Healthy | IPF |
| Test | 2 | 1 |
| Validation | 2 | 1 |
| Training | 12 | 5 |
| Total | 16 | 7 |

## 9.5.1 Multilayer Perceptron Size

We compared combinations of network widths of {100, 200, 300, 400} neurons per layer and network depths of {1, 2, 3, 4} hidden layers. The best network size is shown in Table 9.3.

*Table 9.3: Multilayer perceptron size.*

| Model | Layer | Type | Properties |
|---|---|---|---|
| MLP | 0 | input layer | 4112x1 feature vector |
| | 1 | hidden layer | 300 neurons |
| | 2 | output layer | softmax |

## 9.5.2 Bidirectional Gated Recurrent Neural Network Size

We compared combinations of network widths of {100, 200, 300, 400} neurons per layer and network depths of {1, 2, 3, 4} hidden layers. One half of the number of neurons in each layer is used for the forward layer and the other half for the backward layer. The best network size is shown in Table 9.4.

*Table 9.4: Bidirectional gated recurrent neural network size.*

| Model | Layer | Type | Properties |
|---|---|---|---|
| BiGRNN | 0 | input layer | 4112x1 feature vector |
| | 1 | hidden layer | forward+backward, 200 neurons each |
| | 2 | hidden layer | forward+backward, 200 neurons each |
| | 3 | hidden layer | forward+backward, 200 neurons each |
| | 4 | hidden layer | forward+backward, 200 neurons each |
| | 5 | output | softmax |

### 9.5.3 Convolutional Bidirectional Gated Recurrent Neural Network Size

We used the BiGRNN specified in Table 9.4 and combined it with a CNN front-end to obtain the ConvBiGRNN. We compared combinations of different numbers of feature maps for the first convolutional layer {10, 20, 30, 40, 50, 60, 70, 80, 90} and the second convolutional layer {10, 20, 30, 40, 50, 60, 70, 80, 90}. Details about the architecture are shown in Table 9.5. In the first convolutional layer, we use $1 \times 1$ kernels to reduce the dimension of the input. In the second convolutional layer, we use a $3 \times 3$ kernel with stride and padding of one. No subsampling (pooling) is applied.

*Table 9.5: Convolutional bidirectional gated recurrent neural network size.*

| Model | Layer | Type | Properties |
|---|---|---|---|
| Conv- | 0 | input layer | 4x4x257 image shape (see Figure 9.3) |
| BiGRNN | 1 | convolutional | 30 feature maps, 1x1 kernel |
| | 3 | convolutional | 30 feature maps, 3x3 kernel |
| | 4 | subsampling | not applied |
| | 5 | hidden layer | forward+backward, 200 neurons each |
| | 6 | hidden layer | forward+backward, 200 neurons each |
| | 7 | hidden layer | forward+backward, 200 neurons each |
| | 8 | hidden layer | forward+backward, 200 neurons each |
| | 9 | output | softmax |

### 9.5.4 Comparison of the Three Neural Network Architectures

Table 9.6 shows the results for the different architectures. The reported scores are the micro-average values from the seven folds, evaluated on the test set. The ConvBiGRNN achieves the best results with $F_1 = 92.4\%$.

*Table 9.6: Comparison of different neural networks architectures. Micro-average values from the seven folds evaluated on the test set.*

| Model | $P_+(\%)$ | $Se(\%)$ | $F_1(\%)$ |
|---|---|---|---|
| MLP | 75.0 | 37.8 | 50.2 |
| BiGRNN | 93.1 | 80.0 | 86.1 |
| ConvBiGRNN | **100.0** | **85.9** | **92.4** |

## 9.6 Discussion and Conclusion

In our experiments, we compare different neural network architectures for multi-channel lung sound classification. Firstly, we determine a suitable network size for each architecture using grid search. We compare the architectures of the MLP, the BiGRNN, and the ConvBiGRNN, with the latter outperforming the rest.

As described in Section 5.2.1, adventitious sounds caused by IPF are inspiratory fine (or *velcro*) crackles heard over affected areas [38,86]. Because adventitious sounds are superimposed on normal lung sounds, healthy and pathological recordings mainly differ during inspiration and no distinct difference during expiration can be observed. Within our classification framework, this renders it quite challenging for the MLP, because each frame is classified independent from neighbouring ones. From the three models, the MLP shows the worst performance in terms of F-score (see Table 9.6). Furthermore, it is notable that the Sensitivity is very low. The BiGRNN and the ConvBiGRNN show distinctly better performance. Due to the stacking of the feature vectors of the individual channels within the *basic processing framework*, the dimension $D$ of the resulting feature vector is relatively high. The convolutional front-end reduces the dimension of the feature vectors of the individual channels by using $1 \times 1$ kernels in the first layer. The ConvBiGRNN is able to outperform the BiGRNN with $F_1 = 92.4\%$. It achieves a Precision of $P_+ = 100.0\%$, meaning that all recordings labeled as pathological are actually recognised. The Sensitivity $Se = 85.9\%$ is in an acceptable range.

From a medical point of view, we present an approach for the diagnostic analysis of IPF. Crackles are not specific for IPF, they can also be heard in healthy subjects and can be associated with other diseases, such as congestive heart failure (CHF), COPD, bronchiectasis, and pneumonia [85]. Still, there are notable differences in the temporal occurrence and the characteristics of the crackles. Crackles in IPF appear only during inspiration. Compared to the fine crackles of IPF, those related to CHF and pneumonia are higher in frequency. Another difference, observed by the authors of [165], is that crackles related to IPF are transmitted over a smaller area of the chest than those of CHF and pneumonia. Therefore, multi-channel lung sound analysis provides useful information. Because a misinterpretation of crackles could lead to inappropriate therapy [165], further experiments, including the mentioned diseases, are needed to evaluate if an accurate distinction is possible. Furthermore, the inclusion of metadata should be considered for the classification.

The proposed system enables to decide whether IPF is present or not, but no information about affected areas is provided. The event detection approach presented in Chapter 8 provides information about the temporal occurrence of crackles in the individual sensors and the affected areas. Similar to auscultation by a physician, this can be considered as an intermediate step to detect markers related to the disease. This could be implemented in a two-stage-system: First, adventitious sounds are detected in each channel and, based on the occurring events, the multi-channel recording is classified as normal or pathological. The advantage of the processing framework presented in Section 9.2 is that the complete information from the breathing cycle is taken into account, including the multi-channel information. More information than just the presence or absence of crackles is used. This makes the system more robust against bowel sounds, heart sounds, artefacts, non pathological crackles, and missing or detached sensors. Another advantage is the applicability to diagnose any other lung disease that causes characteristic changes in lung sounds by simply providing appropriate training material.

The classification framework has to cope with several challenges. For thin and/or small subjects, the lowest sensor row and outer sensors of the MLSRD can become irrelevant. For a short torso, the sensors from the lowest row are located over the abdomen and contain mainly bowel sounds. For thin subjects (BMI<20), it is possible that outer sensors are detached. For both cases, we make no distinction in the processing compared to recordings with all sensors fully attached over the lung. Other challenges are the presence of crackles in lung healthy subjects and in general the presence of bowel sound, which mask the lung sounds. Low frequency noise and heart sounds are, to a great extent, already filtered out in the recording stage with the analogue high-pass filter with a cut-off frequency of $f_c = 80\,\mathrm{Hz}$.

As already stated by the authors in [13], the comparison with other attempts to classify lung sounds is difficult, due to the differences in investigated pulmonary pathology, the type of classification scheme, and the data acquisition. Regarding all the mentioned aspects, their work represents the most similar one from literature. As initially mentioned, they use a 25-channel lung sound recording device with a $5 \times 5$ sensor arrangement over the posterior chest. They assess different parameterization techniques for multi-channel lung sounds, which are applied to construct feature vectors used as input to a neural network. For binary classification of healthy vs. interstitial lung diseases (comprising idiopathic pulmonary fibrosis), the parameterization with the univariate autoregressive model results in a classification accuracy of 75% and 93% for healthy subjects and patients with interstitial lung diseases, respectively. A three-class structure for multi-channel lung sound classification is presented in [160]. The three classes are healthy, obstructive, and restrictive, with the latter referring to interstitial lung diseases. The authors model 14-channel lung sounds using a second order 250-point VAR model, and feed the estimated model parameter to SVM and GMM classifiers with various classifier configurations. A hierarchical GMM classifier, which first performs a classification of healthy vs. pathological and subsequently, in the pathological class, of obstructive vs. restrictive, achieves a classification rate of 85%. In the first stage, the sensitivity and the specificity are both 90%.

One limitation of our study is the small number of pathological subjects used in the experiments. We addressed this problem by using cross-validation. Another limitation is the age difference between healthy and pathological subjects, i.e. healthy subjects are much younger than the pathological ones. However, according to [79], the subject's age is not relevant for automated lung auscultation. Although the authors observe age-determined changes of normal lung sounds, these are too small to be clinically relevant. Therefore, the lack of healthy elderly subjects in our database should not be too relevant for the validity of our results.

The proposed method represents a general solution for multi-channel lung sound classification. To evaluate the full potential of our proposed method, further experiments are needed. Therefore, data from various obstructive and restrictive lung disease has to be recorded, and a multiclass classification problem has to be solved.

# 10 Conclusion

In this thesis, we present a holistic approach to multi-channel lung sound analysis. We started from scratch with the development of a multi-channel lung sound recording device. Within a clinical trial, we recorded a lung sound database for lung-healthy subjects and patients diagnosed with idiopathic pulmonary fibrosis (IPF). In terms of lung sound classification, we present several approaches with deep neural networks, applied for event detection and direct diagnosis of the underlying disease. In this context, we also made an excursion to heart sound segmentation.

With respect to recording hardware, we present a 16-channel solution for airflow-aware lung sound analysis. The device records lung sounds over the posterior chest in supine position. Due to the lung sound transducer (LST) design and the attachment method, i.e. the auscultation pad, we achieve robustness in terms of air- and body-borne noise. This allows its usage in real clinical settings, i.e. there is no need for a controlled recording environment. Additionally, the high signal-to-noise ratio of the LST microphone extends the considered frequency range for normal lung sounds towards higher frequency components.

In terms of a multi-channel lung sound database, we present a clinical trial design for data recording and the resulting database. In particular, we recorded lung sounds from 16 lung-healthy subjects and 7 patients diagnosed with IPF at several airflow rates. The recording device was successfully applied for data collection in real clinical settings and the measurement procedure was accepted by the clinical trial subjects.

With regard to heart sound segmentation, we present an event detection approach with deep recurrent neural networks to detect the state sequence *S1 - systole - S2 - diastole*. Compared to the logistic regression hidden semi-Markov model (LR-HSMM)-method, we do not directly incorporate a priori information about the state durations, because the model is capable of learning the temporal dependencies itself. This renders it also applicable to recordings with cardiac arrhythmia and extendable to the detection of extra heart sounds (third and fourth heart sound), heart murmurs, as well as other acoustic events. In experiments, we show state-of-the-art performance on the 2016 PhysioNet/CinC Challenge dataset.

In terms of event detection in lung sounds, we adapted the heart sound segmentation framework to detect crackles and breathing phase events in single-channel lung sound recordings. Although the framework is limited to single-channel processing, in combination with our multi-channel recordings, useful spatial information is provided, i.e. affected areas are identified by the presence of adventitious sounds in specific sensors and their location. In our experiments, we demonstrate this for recordings related to IPF.

For multi-channel classification, we present an approach to exploit spectral, temporal, and spatial information, which also renders exact recording positions unnecessary. In particular, we propose a frame-wise classification framework to process full breathing cycles with convolutional recurrent neural networks. In contrast to our event detection approach, a direct diagnosis of the underlying disease is provided. In our experiments, we achieve $F_1 \approx 92\%$ for the diagnosis/classification of IPF. The system shows robustness against bowel sounds, heart sounds, artefacts, non pathological crackles, and missing or detached sensors.

With our joint consideration of hardware, computational methods, and clinical evaluation issues, we present a significant step forward in computational lung sound analysis (CLSA). Focusing on all three aspects avoids the shift of deficiencies from one to the other domains. The main characteristics and advantages of our approach can be summarised as follows:

- Fast measurement procedure.
- Robust recording stage to eliminate the need for denoising or signal enhancement.
- Robust classification algorithms with no need for exact recording positions.
- General solution, i.e. our approach is applicable for the detection of all kinds of adventitious sounds and for the diagnosis of several lung diseases, given appropriate training material.

## Future Work

Our holistic approach leaves room for improvement and many research questions unanswered. Some are listed in the following:

- *Multi-channel lung sound recording device*:
    - Design of customised hardware for preamplification and analog-to-digital conversion of the LST signal.
- *Multi-channel Lung Sound Database*:
    - Data collection within a large-scale clinical trial with many subjects.
    - Recording of data for various obstructive and restrictive lung diseases.
- *Multi-channel Lung Sound Classification*:
    - Consideration of metadata of the subjects in addition to lung sound recordings.
    - Experiments to discriminate diseases causing similar adventitious sounds, e.g. for IPF, pneumonia, and congestive heart failure (CHF).
    - Extension of the single-channel event detection framework to multi-channel processing.
    - Experiments with other neural network architectures.
    - Investigation of unsupervised machine learning algorithms.
    - Comparison of various feature types.

# 11 List of Publications

[O1] E. Messner, M. Hagmüller, P. Swatek, and F. Pernkopf, "A robust multichannel lung sound recording device," in *Proceedings of the 9th Annual International Conference on Biomedical Electronics and Devices (BIODEVICES'16)*, 2016, pp. 34–39.

[O2] E. Messner, M. Hagmüller, P. Swatek, and F. Pernkopf, "Impact of airflow rate on amplitude and regional distribution of normal lung sounds," in *Proceedings of the 10th Annual International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS'17)*, 2017, pp. 49–53.

[O3] E. Messner, M. Hagmüller, P. Swatek, F.-M. Smolle-Jüttner, and F. Pernkopf, "Respiratory airflow estimation from lung sounds based on regression," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17)*, 2017, pp. 1123–1127.

[O4] E. Messner, M. Zöhrer, and F. Pernkopf, "Heart sound segmentation-an event detection approach using deep recurrent neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1964–1974, 2018.

[O5] E. Messner, M. Fediuk, P. Swatek, S. S., F.-M. Smolle-Jüttner, H. Olschewski, and F. Pernkopf, "Crackle and breathing phase detection in lung sounds with deep bidirectional gated recurrent neural networks," in *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'18)*, 2018, pp. 356–359.

[O6] E. Messner, M. Fediuk, P. Swatek, S. S., F.-M. Smolle-Jüttner, H. Olschewski, and F. Pernkopf, "Multi-channel lung sound classification with convolutional recurrent neural networks," *IEEE Journal of Biomedical and Health Informatics*, 2019 (submitted).

# Bibliography

[1] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PloS one*, vol. 12, no. 5, pp. 1–43, 2017.

[2] S. Reichert, R. Gass, C. Brandt, and E. Andrès, "Analysis of respiratory sounds: state of the art," *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, vol. 2, p. 45, 2008.

[3] A. Gurung, C. G. Scrafford, J. M. Tielsch, O. S. Levine, and W. Checkley, "Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis," *Respiratory medicine*, vol. 105, no. 9, pp. 1396–1403, 2011.

[4] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: a systematic review," *Biocybernetics and Biomedical Engineering*, vol. 33, no. 3, pp. 129–135, 2013.

[5] M. C. Grenier, K. Gagnon, J. J. Genest, J. Durand, and L.-G. Durand, "Clinical comparison of acoustic and electronic stethoscopes and design of a new electronic stethoscope." *The American journal of cardiology*, vol. 81, no. 5, pp. 653–656, 1998.

[6] Z. Dokur, "Respiratory sound classification by using an incremental supervised neural network," *Pattern Analysis and Applications*, vol. 12, no. 4, pp. 309–319, 2009.

[7] S. Matsutake, M. Yamashita, and S. Matsunaga, "Abnormal-respiration detection by considering correlation of observation of adventitious sounds," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 634–638.

[8] N. Nakamura, M. Yamashita, and S. Matsunaga, "Detection of patients considering observation frequency of continuous and discontinuous adventitious sounds in lung sounds," in *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'16)*. IEEE, 2016, pp. 3457–3460.

[9] H. Pasterkamp, S. Kraman, P. DeFrain, and G. Wodicka, "Measurement of respiratory acoustical signals. Comparison of sensors." *CHEST Journal*, vol. 104, no. 5, pp. 1518–1525, 1993.

[10] S. S. Kraman, G. R. Wodicka, G. A. Pressler, and H. Pasterkamp, "Comparison of lung sound transducers using a bioacoustic transducer testing system," *Journal of Applied Physiology*, vol. 101, no. 2, pp. 469–476, 2006.

[11] R. Murphy, "Development of acoustic instruments for diagnosis and management of medical conditions," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 26, no. 1, pp. 16–19, 2007.

[12] I. Sen and Y. Kahya, "A multi-channel device for respiratory sound data acquisition and transient detection," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'06)*, 2006, pp. 6658–6661.

[13] S. Charleston-Villalobos, G. Martinez-Hernandez, R. Gonzalez-Camarena, G. Chi-Lem, J. G. Carrillo, and T. Aljama-Corrales, “Assessment of multichannel lung sounds parameterization for two-class classification in interstitial lung disease patients,” *Computers in biology and medicine*, vol. 41, no. 7, pp. 473–482, 2011.

[14] D. Owens, “Rale lung sounds 3.0,” *CIN: Computers, Informatics, Nursing*, vol. 5, no. 3, pp. 9–10, 2002.

[15] “Listen to lung sounds,” http://www.littmann.ca/wps/portal/3M/en_CA/3M-Littmann-CA/stethoscope/littmann-learning-institute/heart-lung-sounds/lung-sounds/.

[16] S. Lehrer, *Understanding lung sounds, the 2nd edition.* New York, NY, USA: WB Saunders Company, 1993.

[17] ——, *Understanding lung sounds, the 3nd edition.* Philadelphia, PA, USA: WB Saunders Company, 2002.

[18] R. L. Wilkins, J. E. Hodgkin, and B. Lopez, *Fundamentals of lung and heart sounds.* Mosby, 2004.

[19] B. 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.* Springer, 2018, pp. 33–37.

[20] P. Bokov, B. Mahut, P. Flaud, and C. Delclaux, “Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population,” *Computers in biology and medicine*, vol. 70, pp. 40–50, 2016.

[21] D. Chamberlain, R. Kodgule, D. Ganelin, V. Miglani, and R. R. Fletcher, “Application of semi-supervised deep learning to lung sound analysis,” in *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS’16).* IEEE, 2016, pp. 804–807.

[22] J.-C. Chien, H.-D. Wu, F.-C. Chong, and C.-I. Li, “Wheeze detection using cepstral analysis in gaussian mixture models,” in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS’07).* IEEE, 2007, pp. 3168–3171.

[23] B.-S. Lin and B.-S. Lin, “Automatic wheezing detection using speech recognition technique,” *Journal of Medical and Biological Engineering*, vol. 36, no. 4, pp. 545–554, 2016.

[24] E. Ç. Güler, B. Sankur, Y. P. Kahya, and S. Raudys, “Two-stage classification of respiratory sound patterns,” *Computers in biology and medicine*, vol. 35, no. 1, pp. 67–83, 2005.

[25] A. D. Orjuela-Cañón, D. F. Gómez-Cajas, and R. Jiménez-Moreno, "Artificial neural networks for acoustic lung signals classification," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2014, pp. 214–221.

[26] L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, I. Chouvarda, N. Maglaveras, V. Tsara, C. Teixeira, P. Carvalho, J. Henriques *et al.*, "Detection of wheezes using their signature in the spectrogram space and musical features," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'15)*. IEEE, 2015, pp. 5581–5584.

[27] R. Naves, B. H. Barbosa, and D. D. Ferreira, "Classification of lung sounds using higher-order statistics: A divide-and-conquer approach," *Computer methods and programs in biomedicine*, vol. 129, pp. 12–20, 2016.

[28] F. Jin, S. Krishnan, and F. Sattar, "Adventitious sounds identification and extraction using temporal–spectral dominance-based features," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 11, pp. 3078–3087, 2011.

[29] L. Mendes, I. M. Vogiatzis, E. Perantoni, E. Kaimakamis, I. Chouvarda, N. Maglaveras, J. Henriques, P. Carvalho, and R. P. Paiva, "Detection of crackle events using a multi-feature approach," in *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'16)*. IEEE, 2016, pp. 3679–3683.

[30] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2013, pp. 273–278.

[31] M. Zöhrer and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification and acoustic event detection," *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events*, 2016.

[32] M. B. Khodabakhshi and M. H. Moradi, "The attractor recurrent neural network based on fuzzy functions: An effective model for the classification of lung abnormalities," *Computers in biology and medicine*, vol. 84, pp. 124–136, 2017.

[33] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 65, 2017.

[34] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artificial intelligence in medicine*, 2018.

[35] R. Palaniappan, K. Sundaraj, N. U. Ahamed, A. Arjunan, and S. Sundaraj, "Computer-based respiratory sound analysis: a systematic review," *IETE Technical Review*, vol. 30, no. 3, pp. 248–256, 2013.

[36] A. Rizal, R. Hidayat, and H. A. Nugroho, "Signal domain in respiratory sound analysis: Methods, application and future development," *Journal of Computer Science*, vol. 11, no. 10, p. 1005, 2015.

[37] J. Tu, K. Inthavong, and G. Ahmadi, *Computational fluid and particle dynamics in the human respiratory system*. Springer Science & Business Media, 2012.

[38] A. Bohadana, G. Izbicki, and S. S. Kraman, "Fundamentals of lung auscultation," *New England Journal of Medicine*, vol. 370, no. 8, pp. 744–751, 2014.

[39] W. Commons, "File:respiratory system complete en.svg — wikimedia commons, the free media repository," 2018, [Online; accessed 7-January-2019]. [Online]. Available: https://commons.wikimedia.org/w/index.php?title=File:Respiratory_system_complete_en.svg&oldid=292419780

[40] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," *Annals of thoracic medicine*, vol. 10, no. 3, p. 158, 2015.

[41] N. Gavriely, M. Nissan, A. Rubin, and D. W. Cugell, "Spectral characteristics of chest wall breath sounds in normal subjects." *Thorax*, vol. 50, no. 12, pp. 1292–1300, 1995.

[42] E. B. Weiss and C. J. Carlson, "Recording of breath sounds," *American Review of Respiratory Disease*, vol. 105, no. 5, pp. 835–839, 1972.

[43] N. Gavriely, Y. Palti, and G. Alroy, "Spectral characteristics of normal breath sounds," *Journal of Applied Physiology*, vol. 50, no. 2, pp. 307–314, 1981.

[44] P. Forgacs, A. Nathoo, and H. Richardson, "Breath sounds," *Thorax*, vol. 26, no. 3, pp. 288–295, 1971.

[45] American Thoracic Society and others, "Updated nomenclature for membership reaction," *ATS NEWS*, vol. 3, pp. 5–6, 1977.

[46] H. Pasterkamp, S. S. Kraman, and G. R. Wodicka, "Respiratory sounds: advances beyond the stethoscope," *American journal of respiratory and critical care medicine*, vol. 156, no. 3, pp. 974–987, 1997.

[47] N. Gavriely, "Automatic detection and analysis of breath sounds," *European Patent, EP 0*, vol. 951, no. 867, p. A2, 1999.

[48] Y. Nagasaka, "Lung sounds in bronchial asthma," *Allergology International*, vol. 61, no. 3, pp. 353–363, 2012.

[49] N. Meslier, G. Charbonneau, and J. Racineux, "Wheezes," *European respiratory journal*, vol. 8, no. 11, pp. 1942–1948, 1995.

[50] R. P. Baughman and R. G. Loudon, "Stridor: Differentiation from asthma or upper airway noise1-3," *American Review of Respiratory Disease*, vol. 139, pp. 1407–1409, 1989.

[51] N. R. Chamberlain, "Respiratory airway infections," 2014, [Online; accessed 25-January-2019]. [Online]. Available: https://www.atsu.edu/faculty/chamberlain/website/lectures/lecture/reairin2.htm

[52] P. Forgacs, *Lung sounds.* Baillière Tindall, 1978.

[53] R. Paciej, A. Vyshedskiy, D. Bana, and R. Murphy, "Squawks in pneumonia," *Thorax*, vol. 59, no. 2, pp. 177–178, 2004.

[54] M. Munakata, H. Ukita, I. Doi, Y. Ohtsuka, Y. Masaki, Y. Homma, and Y. Kawakami, "Spectral and waveform characteristics of fine and coarse crackles." *Thorax*, vol. 46, no. 9, pp. 651–657, 1991.

[55] A. Jones, "A brief overview of the analysis of lung sounds," *Physiotherapy*, vol. 81, no. 1, pp. 37–42, 1995.

[56] A. Sovijarvi, F. Dalmasso, J. Vanderschoot, L. Malmberg, G. Righini, and S. Stoneman, "Definition of terms for applications of respiratory sounds," *European Respiratory Review*, vol. 10, no. 77, pp. 597–610, 2000.

[57] E. Messner, M. Hagmüller, P. Swatek, and F. Pernkopf, "A robust multichannel lung sound recording device," in *Proceedings of the 9th Annual International Conference on Biomedical Electronics and Devices (BIODEVICES'16)*, 2016, pp. 34–39.

[58] M. Zanartu, J. C. Ho, S. S. Kraman, H. Pasterkamp, J. E. Huber, and G. R. Wodicka, "Airborne and tissue-borne sensitivities of bioacoustic sensors used on the skin surface," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 443–451, 2009.

[59] H. Pasterkamp, G. Wodicka, and S. Kraman, "Effect of ambient respiratory noise on the measurement of lung sounds," *Medical & Biological Engineering & Computing*, vol. 37, no. 4, pp. 461–465, 1999.

[60] S. Liu, R. X. Gao, D. John, J. Staudenmayer, and P. Freedson, "Tissue artifact removal from respiratory signals based on empirical mode decomposition," *Annals of biomedical engineering*, vol. 41, no. 5, pp. 1003–1015, 2013.

[61] A. Sovijarvi, L. Malmberg, G. Charbonneau, J. Vanderschoot, F. Dalmasso, C. Sacco, M. Rossi, and J. Earis, "Characteristics of breath sounds and adventitious respiratory sounds," *European Respiratory Review*, vol. 10, no. 77, pp. 591–596, 2000.

[62] N. Hayashi, "Detection of pneumothorax visualized by computer analysis of bilateral respiratory sounds," *Yonago acta medica*, vol. 54, no. 4, p. 75, 2011.

[63] G. R. Wodicka, S. S. Kraman, G. M. Zenk, and H. Pasterkamp, "Measurement of respiratory acoustic signals. Effect of microphone air cavity depth," *CHEST Journal*, vol. 106, no. 4, pp. 1140–1144, 1994.

[64] S. S. Kraman, G. R. Wodicka, Y. Oh, and H. Pasterkamp, “Measurement of respiratory acoustic signals. Effect of microphone air cavity width, shape, and venting,” *CHEST Journal*, vol. 108, no. 4, pp. 1004–1008, 1995.

[65] L. Vannuccini, J. Earis, P. Helisto, B. Cheetham, M. Rossi, A. Sovijarvi, and J. Vanderschoot, “Capturing and preprocessing of respiratory sounds,” *European Respiratory Review*, vol. 10, no. 77, pp. 616–620, 2000.

[66] “Playrec,” http://www.playrec.co.uk/, accessed: 2014-06-06.

[67] M. Rossi, A. Sovijarvi, P. Piirila, L. Vannuccini, F. Dalmasso, and J. Vanderschoot, “Environmental and subject conditions and breathing manoeuvres for respiratory sound recordings,” *European Respiratory Review*, vol. 10, no. 77, pp. 611–615, 2000.

[68] E. Messner, M. Hagmüller, P. Swatek, and F. Pernkopf, “Impact of airflow rate on amplitude and regional distribution of normal lung sounds,” in *Proceedings of the 10th Annual International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS’17)*, 2017, pp. 49–53.

[69] E. Messner, M. Hagmüller, P. Swatek, F.-M. Smolle-Jüttner, and F. Pernkopf, “Respiratory airflow estimation from lung sounds based on regression,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’17)*. IEEE, 2017, pp. 1123–1127.

[70] S. Kraman, “The relationship between airflow and lung sound amplitude in normal subjects.” *CHEST Journal*, vol. 86, no. 2, pp. 225–229, 1984.

[71] N. Gavriely and D. W. Cugell, “Airflow effects on amplitude and spectral content of normal breath sounds,” *Journal of applied physiology*, vol. 80, no. 1, pp. 5–13, 1996.

[72] I. Hossain and Z. Moussavi, “Relationship between airflow and normal lung sounds,” in *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE’02)*, 2002, pp. 1120–1122.

[73] B. E. Shykoff, Y. Ploysongsang, and H. Chang, “Airflow and normal lung sounds,” *American Review of Respiratory Disease*, vol. 137, pp. 872–876, 1988.

[74] M. Yosef, R. Langer, S. Lev, and Y. A. Glickman, “Effect of airflow rate on vibration response imaging in normal lungs,” *The open respiratory medicine journal*, vol. 3, no. 1, 2009.

[75] A. Torres-Jimenez, S. Charleston-Villalobos, R. Gonzalez-Camarena, G. Chi-Lem, and T. Aljama-Corrales, “Asymmetry in lung sound intensities detected by respiratory acoustic thoracic imaging (RATHI) and clinical pulmonary auscultation,” in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS’08)*, 2008, pp. 4797–4800.

[76] ——, “Respiratory acoustic thoracic imaging (RATHI): Assessing intrasubject variability,” in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS’08)*, 2008, pp. 4793–4796.

[77] S. Charleston-Villalobos, S. Cortés-Rubiano, R. González-Camerena, G. Chi-Lem, and T. Aljama-Corrales, “Respiratory acoustic thoracic imaging (RATHI): assessing deterministic interpolation techniques,” *Medical and Biological Engineering and Computing*, vol. 42, no. 5, pp. 618–626, 2004.

[78] J. A. Fiz, J. Gnitecki, S. S. Kraman, G. R. Wodicka, and H. Pasterkamp, “Effect of body position on lung sounds in healthy young men,” *CHEST Journal*, vol. 133, no. 3, pp. 729–736, 2008.

[79] V. Gross, A. Dittmar, T. Penzel, F. Schuttler, and P. Von Wichert, “The relationship between normal lung sounds, age, and gender,” *American journal of respiratory and critical care medicine*, vol. 162, no. 3, pp. 905–909, 2000.

[80] A. Oliveira and A. Marques, “Respiratory sounds in healthy people: a systematic review,” *Respiratory medicine*, vol. 108, no. 4, pp. 550–570, 2014.

[81] V. Gross, U. Koehler, T. Penzel, C. Reinke, P. Wichert, and C. Vogelmeier, “Der Einfluß der subkutanen Fettschicht auf die normalen Atemgeräusche. The Influence of the Subcutaneous Fat Layer on Normal Lung Sounds,” *Biomedizinische Technik/Biomedical Engineering*, vol. 48, no. 6, pp. 182–185, 2003.

[82] V. Cottin and L. Richeldi, “Neglected evidence in idiopathic pulmonary fibrosis and the importance of early diagnosis and treatment,” *European Respiratory Review*, vol. 23, no. 131, pp. 106–110, 2014.

[83] D. J. Lamas, S. M. Kawut, E. Bagiella, N. Philip, S. M. Arcasoy, and D. J. Lederer, “Delayed Access and Survival in Idiopathic Pulmonary Fibrosis: A Cohort Study,” *American Journal of Respiratory and Critical Care Medicine*, vol. 187, no. 7, pp. 842–847, 2011.

[84] B. Ley and H. Collard, “Epidemiology of idiopathic pulmonary fibrosis,” *Clinical Epidemiology*, vol. 5, pp. 483–492, 2013.

[85] V. Cottin and J.-F. Cordier, “Velcro crackles: the key for early diagnosis of idiopathic pulmonary fibrosis?” *European Respiratory Journal*, vol. 40, no. 3, pp. 519–521, 2012.

[86] B. Flietstra, N. Markuzon, A. Vyshedskiy, and R. Murphy, “Automated analysis of crackles in patients with interstitial pulmonary fibrosis,” *Pulmonary medicine*, vol. 2011, 2011.

[87] Ş. Kaya, A. A. Çevik, N. Acar, E. Döner, C. Sivrikoz, and R. Özkan, “A study on the evaluation of pneumothorax by imaging methods in patients presenting to the emergency department for blunt thoracic trauma,” *Turkish Journal of Trauma and Emergency Surgery*, vol. 21, no. 5, pp. 366–372, 2015.

[88] A. Lamb, M. Qadan, and A. Gray, “Detection of occult pneumothoraces in the significantly injured adult with blunt trauma,” *European Journal of Emergency Medicine*, vol. 14, no. 2, pp. 65–67, 2007.

[89] S. Ramakrishnan, S. Udpa, and L. Udpa, “A numerical model to study auscultation sounds under pneumothorax conditions,” in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS’09)*, 2009, pp. 6201–6204.

[90] E. Messner, M. Fediuk, P. Swatek, S. S., F.-M. Smolle-Jüttner, H. Olschewski, and F. Pernkopf, “Multi-channel lung sound classification with convolutional recurrent neural networks,” *IEEE Journal of Biomedical and Health Informatics*, 2019 (submitted).

[91] E. Messner, M. Zöhrer, and F. Pernkopf, “Heart sound segmentation—an event detection approach using deep recurrent neural networks,” *IEEE transactions on biomedical engineering*, vol. 65, no. 9, pp. 1964–1974, 2018.

[92] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[93] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Neurocomputing: Foundations of research,” J. A. Anderson and E. Rosenfeld, Eds. Cambridge, MA, USA: MIT Press, 1988, ch. Learning Internal Representations by Error Propagation, pp. 673–695.

[94] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[95] A. Graves, A. rahman Mohamed, and G. E. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’13)*, 2013, pp. 6645–6649.

[96] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.

[97] ——, “Gated feedback recurrent neural networks,” in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2067–2075.

[98] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *CoRR*, vol. abs/1412.3555, 2014.

[99] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[100] Y. LeCun, Y. Bengio *et al.*, “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[101] T. Miyato, S. Shin-ichi Maeda, M. Koyama, K. Nakae, and S. Ishii, “Distributional smoothing by virtual adversarial examples.” *CoRR*, vol. abs/1507.00677, 2015.

[102] M. Ratajczak, S. Tschiatschek, and F. Pernkopf, "Virtual adversarial training applied to neural higher-order factors for phone classification." in *Interspeech*, 2016, pp. 2756–2760.

[103] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow, "Adversarial autoencoders," *CoRR*, vol. abs/1511.05644, 2015.

[104] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[105] 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.

[106] B. Poole, J. Sohl-Dickstein, and S. Ganguli, "Analyzing noise in autoencoders and deep networks," *arXiv preprint arXiv:1406.1831*, 2014.

[107] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. Johnson *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, p. 2181, 2016.

[108] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *Computers in Cardiology, 1997*. IEEE, 1997, pp. 105–108.

[109] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on S-transform," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 273–281, 2013.

[110] S. Sun, Z. Jiang, H. Wang, and Y. Fang, "Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform," *Computer Methods and Programs in Biomedicine*, vol. 114, no. 3, pp. 219–230, 2014.

[111] S. Choi and Z. Jiang, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1056–1069, 2008.

[112] Z. Yan, Z. Jiang, A. Miyamoto, and Y. Wei, "The moment segmentation analysis of heart sound pattern," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 2, pp. 140–150, 2010.

[113] S. Ari, P. Kumar, and G. Saha, "A robust heart sound segmentation algorithm for commonly occurring heart valve diseases," *Journal of Medical Engineering & Technology*, vol. 32, no. 6, pp. 456–465, 2008.

[114] H. Naseri and M. Homaeinezhad, "Detection and boundary identification of phonocardiogram sounds using an expert frequency-energy based metric," *Annals of Biomedical Engineering*, vol. 41, no. 2, pp. 279–292, 2013.

[115] D. Kumar, P. Carvalho, M. Antunes, J. Henriques, L. Eugenio, R. Schmidt, and J. Habetha, "Detection of S1 and S2 heart sounds by high frequency signatures," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'06)*. IEEE, 2006, pp. 1410–1416.

[116] V. N. Varghees and K. Ramachandran, "A novel heart sound activity detection framework for automated heart sound analysis," *Biomedical Signal Processing and Control*, vol. 13, pp. 174–188, 2014.

[117] J. Pedrosa, A. Castro, and T. T. Vinhoza, "Automatic heart sound segmentation and murmur detection in pediatric phonocardiograms," in *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'14)*. IEEE, 2014, pp. 2294–2297.

[118] J. Vepa, P. Tolay, and A. Jain, "Segmentation of heart sounds using simplicity features and timing information," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*. IEEE, 2008, pp. 469–472.

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

[120] A. Gharehbaghi, T. Dutoit, A. Sepehri, P. Hult, and P. Ask, "An automatic tool for pediatric heart sounds segmentation," in *Computing in Cardiology, 2011*. IEEE, 2011, pp. 37–40.

[121] T. Oskiper and R. Watrous, "Detection of the first heart sound using a time-delay neural network," in *Computers in Cardiology, 2002*. IEEE, 2002, pp. 537–540.

[122] A. A. Sepehri, A. Gharehbaghi, T. Dutoit, A. Kocharian, and A. Kiani, "A novel method for pediatric heart sound segmentation without using the ECG," *Computer Methods and Programs in Biomedicine*, vol. 99, no. 1, pp. 43–48, 2010.

[123] T. Chen, K. Kuan, L. A. Celi, and G. D. Clifford, "Intelligent heartsound diagnostics on a cellphone using a hands-free kit." in *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.

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

[125] H. Tang, T. Li, T. Qiu, and Y. Park, "Segmentation of heart sounds based on dynamic clustering," *Biomedical Signal Processing and Control*, vol. 7, no. 5, pp. 509–516, 2012.

[126] S. Rajan, E. Budd, M. Stevenson, and R. Doraiswami, "Unsupervised and uncued segmentation of the fundamental heart sounds in phonocardiograms using a time-scale representation," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'06)*. IEEE, 2006, pp. 3732–3735.

[127] T.-E. Chen, S.-I. Yang, L.-T. Ho, K.-H. Tsai, Y.-H. Chen, Y.-F. Chang, Y.-H. Lai, S.-S. Wang, Y. Tsao, and C.-C. Wu, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 372–380, 2017.

[128] L. Gamero and R. Watrous, "Detection of the first and second heart sound using probabilistic models," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'03)*, vol. 3. IEEE, 2003, pp. 2877–2880.

[129] A. D. Ricke, R. J. Povinelli, and M. T. Johnson, "Automatic segmentation of heart sound signals using hidden Markov models," in *Computers in Cardiology, 2005*. IEEE, 2005, pp. 953–956.

[130] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in *Computers in Cardiology, 2005*. IEEE, 2005, pp. 957–960.

[131] P. Sedighian, A. W. Subudhi, F. Scalzo, and S. Asgari, "Pediatric heart sound segmentation using hidden Markov model," in *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'14)*. IEEE, 2014, pp. 5490–5493.

[132] A. Castro, T. T. Vinhoza, S. S. Mattos, and M. T. Coimbra, "Heart sound segmentation of pediatric auscultations using wavelet analysis," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'13)*. IEEE, 2013, pp. 3909–3912.

[133] S. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiological Measurement*, vol. 31, no. 4, p. 513, 2010.

[134] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 822–832, 2016.

[135] C. Liu, D. Springer, and G. D. Clifford, "Performance of an open-source heart sound segmentation algorithm on eight independent databases," *Physiological measurement*, vol. 38, no. 8, p. 1730, 2017.

[136] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *CoRR*, vol. abs/1409.1259, 2014.

[137] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 3104–3112.

[138] T. c. I. Yang and H. Hsieh, "Classification of acoustic physiological signals based on deep learning neural networks with augmented features," in *2016 Computing in Cardiology Conference (CinC)*, Sept 2016, pp. 569–572.

[139] C. Thomae and A. Dominik, “Using deep gated RNN with a convolutional front end for end-to-end classification of heart sound,” in *2016 Computing in Cardiology Conference (CinC)*, Sept 2016, pp. 625–628.

[140] J. van der Westhuizen and J. Lasenby, “Bayesian LSTMs in medicine,” *arXiv preprint arXiv:1706.01242*, 2017.

[141] O. Gencoglu, T. Virtanen, and H. Huttunen, “Recognition of acoustic events using deep neural networks,” in *Proceedings of the 22nd European Signal Processing Conference*, 2014, pp. 506–510.

[142] G. Chen, C. Parada, and T. N. Sainath, “Query-by-example keyword spotting using long short-term memory networks,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’15)*, 2015, pp. 5236–5240.

[143] G. Parascandolo, H. Huttunen, and T. Virtanen, “Recurrent neural networks for polyphonic sound event detection in real life recordings,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’16)*, 2016, pp. 6440–6444.

[144] “Sound exchange,” http://sox.sourceforge.net, accessed: 2017-07-05.

[145] C. Thomae and A. Dominik, “Using deep gated RNN with a convolutional front end for end-to-end classification of heart sound,” in *Computing in Cardiology Conference (CinC), 2016*. IEEE, 2016, pp. 625–628.

[146] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[147] A. M. Saxe, J. L. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” *International Conference of Learning Representations (ICLR)*, 2014.

[148] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.

[149] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.

[150] E. Van der Wall and M. Schalij, “Mitral valve prolapse: a source of arrhythmias?” *The international journal of cardiovascular imaging*, vol. 26, no. 2, pp. 147–149, 2010.

[151] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F.-M. Smolle-Juttner, H. Olschewski, and F. Pernkopf, “Crackle and breathing phase detection in lung sounds with deep bidirectional gated recurrent neural networks,” in *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS’18)*, 2018, pp. 356–359.

[152] A. Sovijärvi, J. Vanderschoot, and J. Earis, "Standardization of computerized respiratory sound analysis," *European Respiratory Review*, vol. 10, no. 77, pp. 974–987, 2000.

[153] S. Huq and Z. Moussavi, "Automatic breath phase detection using only tracheal sounds," in *Proceedings of the 32th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'10)*. IEEE, 2010, pp. 272–275.

[154] Z. K. Moussavi, M. T. Leopando, and G. R. Rempel, "Automated detection of respiratory phases by acoustical means," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'98)*. IEEE, 1998, pp. 21–24.

[155] Z. Moussavi, M. T. Leopando, H. Pasterkamp, and G. Rempel, "Computerised acoustical respiratory phase detection without airflow measurement," *Medical and Biological Engineering and Computing*, vol. 38, no. 2, pp. 198–203, 2000.

[156] X. Lu and M. Bahoura, "An automatic system for crackles detection and classification," in *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'06)*, 2006, pp. 725–729.

[157] C. Pinho, A. Oliveira, C. Jácome, J. Rodrigues, and A. Marques, "Automatic crackle detection algorithm based on fractal dimension and box filtering," *Procedia Computer Science*, vol. 64, pp. 705–712, 2015.

[158] G. Serbes, C. O. Sakar, Y. P. Kahya, and N. Aydin, "Pulmonary crackle detection using time–frequency and time–scale analysis," *Digital Signal Processing*, vol. 23, no. 3, pp. 1012–1021, 2013.

[159] R. Murphy, "Computerized multichannel lung sound analysis," *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, no. 1, p. 16, 2007.

[160] I. Sen, M. Saraclar, and Y. P. Kahya, "A comparison of SVM and GMM-based classifier configurations for diagnostic classification of pulmonary sounds," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 7, pp. 1768–1776, 2015.

[161] G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen *et al.*, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[162] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3461–3466.

[163] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*, March 2016, pp. 5200–5204.

[164] M. Lin, Q. Chen, and S. Yan, “Network in network,” *arXiv preprint arXiv:1312.4400*, 2013.

[165] A. Vyshedskiy, F. Bezares, R. Paciej, M. Ebril, J. Shane, and R. Murphy, “Transmission of crackles in patients with interstitial pulmonary fibrosis, congestive heart failure, and pneumonia,” *CHEST Journal*, vol. 128, no. 3, pp. 1468–1474, 2005.

# A

# Abbreviations

| | |
|---|---|
| **AAE** | acoustic airflow estimation |
| **AD** | aortic disease |
| **ADAT** | Alesis Digital Audio Tape |
| **AGES** | Austrian Agency for Health and Food Safety |
| **ANN** | artificial neural network |
| **AS** | aortic stenosis |
| **BiGRNN** | bidirectional gated recurrent neural network |
| **BiLSTM** | bidirectional long short-term memory |
| **BiRNN** | bidirectional recurrent neural network |
| **BMI** | body mass index |
| **BPD** | breathing phase detection |
| **CAD** | coronary artery disease |
| **CAS** | continuous adventitious sounds |
| **CHF** | congestive heart failure |
| **CEE** | cross-entropy error |
| **CinC** | Computing in Cardiology |
| **CLSA** | computational lung sound analysis |
| **ConvBiGRNN** | convolutional bidirectional gated recurrent neural network |
| **COPD** | chronic obstructive pulmonary disease |
| **CORSA** | computerised respiratory sound analysis |
| **CRF** | case report form |
| **CT** | computed tomography |
| **CUDA** | Compute Unified Device Architecture |
| **CNN** | convolutional neural network |
| **CRNN** | convolutional recurrent neural network |
| **DAS** | discontinuous adventitious sounds |
| **DC** | direct current |
| **DNN** | deep neural network |
| **DRNN** | deep recurrent neural network |
| **ECMC** | electret-condenser microphone capsule |
| **EEC** | European Economic Community |
| **EN** | European Standard |

| | |
|---|---|
| **FNN** | feedforward neural network |
| **GCP** | Good Clinical Practice |
| **GMM** | Gaussian mixture model |
| **GPU** | graphics processing unit |
| **GRNN** | gated recurrent neural network |
| **GRU** | gated recurrent unit |
| **GT** | ground truth |
| **GUI** | graphical user interface |
| **HMM** | hidden Markov model |
| **ICH** | International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use |
| **IPF** | idiopathic pulmonary fibrosis |
| **ISO** | International Organization for Standardization |
| **k-NN** | k-nearest neighbours algorithm |
| **KL** | Kullback–Leibler |
| **LFCC** | linear frequency cepstral coefficient |
| **LR-HSMM** | logistic regression hidden semi-Markov model |
| **LST** | lung sound transducer |
| **LSTM** | long short-term memory |
| **MAR** | multivariate autoregressive |
| **MFCC** | mel frequency cepstral coefficient |
| **MLP** | multilayer perceptron |
| **MLSRD** | multi-channel lung sound recording device |
| **MPC** | miscellaneous pathological condition |
| **MR** | mitral regurgitation |
| **MSE** | mean squared error |
| **MUG** | Medical University of Graz |
| **MVP** | mitral valve prolapse |
| **PCG** | phonocardiogram |
| **PN** | PhysioNet |
| **PSD** | power spectral density |
| **RATHI** | respiratory acoustic thoracic image |
| **RNN** | recurrent neural network |
| **S1** | first heart sound |
| **S2** | second heart sound |
| **S3** | third heart sound |
| **S4** | fourth heart sound |
| **SNR** | signal-to-noise ratio |
| **STFT** | short-time Fourier transform |
| **SVM** | support vector machine |
| **UAR** | univariate autoregressive |

| | |
|---|---|
| **USB** | Universal Serial Bus |
| **VAR** | vector autoregressive |
| **VAT** | virtual adversarial training |

# B
# Symbols

| | |
|---|---|
| $t$ | time index |
| $T$ | number of samples (time domain) |
| $\mathbf{x}_t$ | audio recording in the time domain |
| $l$ | layer index |
| $L$ | number of layers |
| $f$ | frame index (or frequency) |
| $F$ | number of frames |
| $i$ | input gate |
| $u$ | forget gate |
| $r$ | reset gate |
| $o$ | output gate |
| $z$ | update gate |
| $c$ | memory cell |
| $\widetilde{c}$ | new memory cell content |
| $\mathbf{x}^l_f$ | input vector |
| $\mathbf{y}_f$ | output vector |
| $\mathbf{W}^l_x$ | input weight matrix |
| $\mathbf{W}_y$ | output weight matrix |
| $\mathbf{W}^l_h$ | hidden weight matrix |
| $\mathbf{W}^l_{xi}$ | input weight matrix of input gate |
| $\mathbf{W}^l_{hi}$ | hidden weight matrix of input gate |
| $\mathbf{W}^l_{xu}$ | input weight matrix of forget gate |
| $\mathbf{W}^l_{hu}$ | hidden weight matrix of forget gate |
| $\mathbf{W}^l_{xr}$ | input weight matrix of reset gate |
| $\mathbf{W}^l_{hr}$ | hidden weight matrix of reset gate |
| $\mathbf{W}^l_{xo}$ | input weight matrix of output gate |
| $\mathbf{W}^l_{ho}$ | hidden weight matrix of output gate |
| $\mathbf{W}^l_{xc}$ | input weight matrix of memory cell |
| $\mathbf{W}^l_{hc}$ | hidden weight matrix of memory cell |
| $\mathbf{b}^l_h$ | hidden bias vector |
| $\mathbf{b}_y$ | output bias vector |
| $\mathbf{b}^l_u$ | bias vector of forget gate |

| | |
|---|---|
| $\mathbf{b}_r^l$ | bias vector of reset gate |
| $\mathbf{b}_o^l$ | bias vector of output gate |
| $\mathbf{b}_c^l$ | bias vector of memory cell |
| $\mathbf{h}_f^l$ | hidden states vector |
| $\mathbf{i}_f^l$ | input states vector |
| $\mathbf{u}_f^l$ | forget states vector |
| $\mathbf{r}_f^l$ | reset states vector |
| $\mathbf{o}_f^l$ | output states vector |
| $\mathbf{c}_f^l$ | memory cell states vector |
| $\tilde{\mathbf{c}}_f^l$ | new memory states vector |
| $\mathbf{z}_f$ | update states vector |
| $\tilde{\mathbf{h}}_f$ | candidate activation vector |
| $\overrightarrow{\mathbf{h}}_f^l$ | forward hidden sequence states vector |
| $\overleftarrow{\mathbf{h}}_f^l$ | backward hidden sequence states vector |
| $g(\cdot)$ | non-linear function |
| $m(\cdot)$ | non-linear function |
| $\sigma(\cdot)$ | sigmoid function |
| $\odot$ | element-wise product operator |
| $N$ | width and height of a 2-dimensional image |
| $m$ | dimension of kernel |
| $K$ | number of feature maps |
| $k$ | index of feature map |
| $\mathbf{X}_{i,j}^l$ | section of input image |
| $i, j$ | position indexes |
| $\mathbf{W}^{kl}$ | kernel matrix |
| $b_k^l$ | bias term |
| $h_{ij}^{kl}$ | feature map (activation) |
| $n$ | pooling dimension |
| $p(\mathbf{y}_f \mid \mathbf{x}_f)$ | posterior distribution |
| $\boldsymbol{\delta}_f$ | adversarial perturbation |
| $(\cdot\Vert\cdot)$ | Kullback-Leibler (KL)-divergence |
| $\epsilon$ | limit for maximum perturbation |
| $I_p$ | number of iterations |
| $\lambda$ | tradeoff parameter |
| $\mathbf{W}_{x,test}^l$ | scaled weights during testing with dropout |
| $\mathbf{v}^l$ | vector of independent Bernoulli random variables |
| $p$ | probability |
| $\tilde{\mathbf{x}}_f^l$ | thinned input vector |
| $*$ | convolution operator |
| $f_c$ | cut-off frequency |
| $f_s$ | sampling frequency |

| | |
|---|---|
| $f$ | frequency (or frame index) |
| $L_{Aeq}$ | A-weighted equivalent sound level |
| $d$ | distance |
| $f_L$ | lower cut-off frequency |
| $f_H$ | upper cut-off frequency |
| $R^2$ | coefficient of determination |
| $T_w$ | duration of window |
| $M$ | number of filter banks |
| $C$ | number of cepstral coefficients |
| $f_{s,new}$ | new sampling frequency (after downsampling) |
| $D$ | dimension of the feature vector |
| $\mathbb{R}^D$ | set of real numbers |
| $TP$ | true positives |
| $FP$ | false positives |
| $FN$ | false negatives |
| $P_+$ | Precision |
| $Se$ | Sensitivity |
| $F_1$ | F-score |
| $N_{ref}$ | number of reference states |
| $N_{sys}$ | number of system states |
| $N_{TP}$ | number of true positives |
| $N_{FN}$ | number of false negatives |
| $N_{FP}$ | number of false positives |
| $\Delta$ | delta coefficients |
| $\Delta^2$ | acceleration coefficients |
| $\tilde{\mathbf{y}}_f$ | real-valued output vector |
| $\tilde{\mathbf{y}}_{sum}$ | sum over all frames F of real-valued output vector |