BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

DOCTORAL THESIS
SHORTENED VERSION

Brno, 2019 Ing. Pavol Harár

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF TELECOMMUNICATIONS
ÚSTAV TELEKOMUNIKACÍ

AUDIO CLASSIFICATION WITH DEEP LEARNING ON LIMITED DATA SETS
KLASIFIKACE AUDIA HLUBOKÝM UČENÍM S LIMITOVANÝMI ZDROJI DAT

DOCTORAL THESIS
DIZERTAČNÍ PRÁCE

AUTHOR — **Ing. Pavol Harár**
AUTOR PRÁCE

SUPERVISOR — **Ing. Jiří Mekyska, Ph.D.**
ŠKOLITEL

BRNO 2019

ABSTRACT

Standard procedures of dysphonia diagnosis by a clinical speech therapist have their downsides, mainly because the process is very subjective. Recently, an automatic objective analysis of a speaker's condition gained in popularity. Researchers successfully based their methods on various machine learning algorithms and handcrafted features. These methods, unfortunately, are not directly scalable to other voice disorders and the process of feature engineering is laborious and thus financially and talent expensive. Based on the previous successes, a deep learning approach might help to ease the problems with scalability and generalization, but an obstacle is a limited amount of training data. This is a common denominator in almost all systems for automated medical data analysis. The main aim of this work is to research new approaches to deep-learning-based predictive modeling using limited audio data sets, focusing especially on voice pathology assessment. This work is the first to experiment with deep learning in this field and on so far the largest combined database of dysphonic voices, which was created in this work. It provides a thorough examination of publicly available data sources and identifies their limitations. It describes the design of novel time-frequency representations based on Gabor transform and it presents a new class of loss functions, that yield target representations beneficial for learning. In numerical experiments, it demonstrates improvements in the performance of convolutional neural networks trained on limited audio data sets using the augmented target loss function and the newly proposed time-frequency representations, namely Gabor and Mel scattering.

KEYWORDS

deep learning, voice pathologies, Gabor scattering, limited data, audio

ARCHIVED IN

The full version of this dissertation is available at the Science department of the Dean's Office of The Faculty of Electrical Engineering and Communication of Brno University of Technology, Technická 10, Brno, 616 00, Czech Republic.

ABSTRAKT

Standardní postupy diagnózy dysfonie klinickým logopedem mají své nevýhody, především tu, že je tento proces velmi subjektivní. Nicméně v poslední době získala popularitu automatická objektivní analýza stavu mluvčího. Vědci úspěšně založili své metody na různých algoritmech strojového učení a ručně vytvořených příznacích. Tyto metody nejsou bohužel přímo škálovatelné na jiné poruchy hlasu, samotný proces tvorby příznaků je pracný a také náročný z hlediska financí a talentu. Na základě předchozích úspěchů může přístup založený na hlubokém učení pomoci překlenout některé problémy se škálovatelností a generalizací, nicméně překážkou je omezené množství trénovacích dat. Jedná se o společný jmenovatel téměř ve všech systémech pro automatizovanou analýzu medicínských dat. Hlavním cílem této práce je výzkum nových přístupů prediktivního modelování založeného na hlubokém učení využívající omezené sady zvukových dat, se zaměřením zejména na hodnocení patologických hlasů. Tato práce je první, která experimentuje s hlubokým učením v této oblasti, a to na dosud největší kombinované databázi dysfonických hlasů, která byla v rámci této práce vytvořena. Předkládá důkladný průzkum veřejně dostupných zdrojů dat a identifikuje jejich limitace. Popisuje návrh nových časově-frekvenčních reprezentací založených na Gaborově transformaci a představuje novou třídu chybových funkcí, které přinášejí reprezentace výstupů prospěšné pro učení. V numerických experimentech demonstruje zlepšení výkonu konvolučních neuronových sítí trénovaných na omezených zvukových datových sadách pomocí tzv. "augmented target loss function" a navržených časově-frekvenčních reprezentací "Gabor" a "Mel scattering".

KLÍČOVÁ SLOVA

hluboké učení, patologie hlasu, Gabor scattering, limitovaná data, zvuk

MÍSTO ULOŽENÍ PRÁCE

Plná verze této disertační práce je k dispozici na Vědeckém oddělení děkanátu Fakulty Elektrotechniky a Komunikačních Technologií, Vysoké Učení Technické v Brně, Technická 10, Brno, 616 00, Česká Republika.

Contents

Preamble

Publications

Appendix

Preamble

1 Introduction

> *"The potential benefits are huge; everything that civilisation has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools that AI may provide, but the eradication of war, disease, and poverty would be high on anyone's list. Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks."*
>
> *Stephen Hawking, Stuart Russell, Max Tegmark & Frank Wilczek, 2014*

As a society, we should not try to stand in the way of technological advances. I believe, there is no way of stopping what we have already been once able to imagine. However, we also should not completely surrender to anything that just becomes convenient. Instead, we should use the new technology with caution, without locking ourselves and other species on the planet unwarily out of other options. While keeping such an open, but a watchful mindset, we should focus on finding ways to take advantage of our innovations to ease the suffering where we see possible.

During my doctoral studies, I have tried to focus my research work in that manner – to embrace new ideas and further their development towards a wholesome application. It culminates in this thesis in the form of a cumulative dissertation, which comprises a certain portion of the published works produced by my coauthors and me. It gives a brief introduction to each of the relevant topics and describes the genesis of the presented ideas. Furthermore, it provides a story line contextually linking and summarizing the individual papers. Finally, the work as a whole is discussed and concluded.

From a methodological point of view, the central idea of this scientific endeavor is an exploration of the learning capabilities of deep neural networks trained with audio data, particularly in sequence classification. From the application perspective, we explore and address the domain-specific challenges which emerge in the analysis of pathological voices.

This document is structured into three main parts, namely the Preamble, Publications, and Appendix. Those lines describing my ideas and points of view are written in the singular form of the first person, the rest, summarizing the joint effort of my coauthors and me, is written in the plural form of the first person.

In the following sections, I am introducing the relevant topics in a non-technical, colloquial way. The main aim is to provide a potential reader without sufficient background, an idea, where these topics originate. A more experienced reader with an understanding of deep learning, audio signal processing, and medical data analysis, whom this thesis assumes, may consider skipping the Introduction and continuing directly to Summary of the Publications. If this is not the case, I recommend further reading, at the end of each section.

1.1 Deep Learning

Artificial intelligence (AI) is a field of study in computer science (CS). Its definition is unfortunately not clear or straightforward, as it took Stuart Russell and Peter Norvig exactly 31 full pages in their book *Artificial Intelligence: A Modern Approach (2016)* [44], to introduce, define and summarize the concept along with its philosophical, mathematical and other cultural foundations. Colloquially, it refers to the ability of a machine to solve a task by imitating intelligent human behavior. It is often confused with artificial general intelligence (AGI), which inspired a multitude of science fiction authors because of the fascinating idea of a computer that would be equally intelligent to humans in every aspect. The term "Artificial intelligence" was coined by John McCarthy in 1955 [8], just about a year after the sad death of Alan Turing, the father of CS [4].

> *"It would be useful if computers could learn from experience and thus automatically improve the efficiency of their own programs during execution."*
>
> *Donald Michie, 1968*

Machine learning (ML), a subfield of AI, is also a term which refers to a particular set of algorithms, that enable the computers to learn from historical data i.e. experience, without being explicitly programmed. This is a paraphrased quote often attributed to Arthur Samuel, who is also considered having coined the term "Machine learning" back in 1959 [46]. According to Russell and Norvig, ML is a capability of a computer to adapt to new circumstances and to detect and extrapolate patterns [44]. A ML algorithm builds a mathematical model based on the set of training data, which provides an approximation of an unknown optimal solution of the task as measured by a performance metric.

"We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data."

Yann LeCun, Yoshua Bengio & Geoffrey Hinton, 2015

Deep learning (DL) is a subfield of ML concerned with artificial neural networks (ANN). ANNs are computation systems of interconnected artificial neurons, which very loosely model the biological neurons. ANNs have been developed since 1943, when McCulloch and Pitts, inspired by the study of the human brain modeled an electrical circuit of a simple neural network [35], and since Rosenblatt described a mathematical model of Perceptron in 1958 [43]. Nowadays, ANNs are usually described as directed graphs of nodes connected with edges and organized into layers.

The term "Deep learning" was introduced by Rina Dechter in 1986 [47]. The word "deep" refers to a subset of ANNs with a number of hidden layers (number of layers excluding input and output layer) bigger than one. Depending on how the nodes are linked, i.e. the topology, the deep neural network (DNN) is either feedforward or recurrent. Edges represent weights, which parameterize the model and are adjusted during the training of the network.

DNNs were not popular at first but became widely used with the increased availability of data and computing power. In the past years, they dramatically improved the state of the art in areas such as speech recognition, visual object recognition, robotics, bioinformatics, online advertising, search engines, and medical applications [30, 31], to name a few. In this work, we are mainly concerned with architectures composed of one or more of the following components: standard fully connected feedforward layers, convolutional layers as introduced in deep convolutional neural networks (CNN) and recurrent long short-term memory layers (LSTM) [29, 25, 45].

For further reading, please refer to the following books, which go into a great detail in each topic: *Artificial Intelligence: A Modern Approach (Russell & Norvig, 2016)* [44], *Pattern recognition and machine learning (Bishop, 2011)* [5], *Introduction to Machine Learning (Alpaydin, 2014)* [2], *Deep Learning (Goodfellow, Bengio & Courville, 2016)* [17]. For a quicker overview of DL, refer to the works of LeCun, Bengio & Hinton (2015) [30], Schmidhuber (2016) [47], Liu et al. (2017) [32] and Pouyanfar et al. (2018) [41].

1.2 Digital Audio Signal Processing

To process sound information with neural networks, it is necessary to transform the continuous acoustic physical phenomenon into its discrete, digital, computer-understandable representation, i.e. audio data. The field concerned with recording real-world signals like voice, music, etc., their further conversion and processing is called digital signal processing (DSP). In this work, we will be mainly interested in decisions of sampling rate and time-frequency representations of the sound, as well as psychoacoustics and we will study their impact on learning.

For a comprehensive introduction into these topics, please refer to the following books: *Discrete-Time Signal Processing (Oppenheim & Schafer, 2014)* [39], *Digital Audio Signal Processing (Zölzer, 2008)* [48], *Understanding Digital Signal Processing (Lyons, 2004)* [33], *Foundations of Time-Frequency Analysis (Gröchenig, 2001)* [18]. For a more concise merger introducing DL from the perspective of audio signal processing, refer to the paper by Purwins et al. (2019) [42].

1.3 Automatic Analysis of Medical Audio Data

According to an extensive survey in medical image analysis by Litjens et al. (2017) [31], medical images have been automatically analyzed as soon as it was possible to capture and load them into a computer. In the case of audio, researchers were first interested in using extralinguistic information to identify speakers, their age or gender. For speech emotion recognition, they have used paralinguistic information, and in the case of accent, dialect or speech recognition, the linguistic dimension has been studied. Just in the past years, the analysis of the speaker's condition gained in popularity, as Gómez-García, Moro-Velázquez & Godino-Llorente (2019) [16] explain in another great survey on automatic voice condition analysis (AVCA) systems. AVCA aims for an objective and automatic quantification of the degree to which a patient is impaired by a voice disorder. One of the main advantages of such analysis based on audio data is its relatively low cost, non-invasive nature and a possibility for continuous monitoring and in-cloud processing [36].

The fact that sparked my interest in this research direction was a link between hypokinetic dysarthria (HD) and Parkinson's disease (PD). HD is a motor speech disorder manifested in articulation, phonation, prosody, respiration, and faciokinesis, that occurs in up to 90 % of PD patients and is also considered one of the early markers of PD. For more information about HD and other disorders in PD, please refer to a thorough survey paper by Brabenec et al. (2017) [6]. Unfortunately, nowadays, it is still not possible to cure PD, but an early diagnosis can significantly improve patient's quality of life thanks to already available medication.

The standard procedure of HD diagnosis is carried out by a clinical speech therapist. Speech and voice of a patient are usually assessed using specific scales and questionnaires such as Frenchay dysarthria assessment [12] or 3F test [27]. This procedure still has its downsides though, mainly because the evaluations are very subjective. The human ear, even of a trained clinician, is not sensitive enough to capture slight changes in the patient's voice or speech, it is, therefore, hard to compare successive assessments for progression tracking, even from the same clinician [36].

Researchers thus started to work on automatic objective methods of HD analysis and proposed a variety of parameterization methods, to extract conventional or non-conventional features from the audio recordings of the patients' speech and voice. These were further utilized in predictive modeling using a variety of machine learning techniques [6] to infer an automatic evaluation of the patient's data. Successes of these methods are undeniable and encouraging, with strong advantages for clinicians who can use these methods as a supportive tool for their decisions. The main pros are objectivity and relatively good interpretability [11], which is important in this setting. Unfortunately, the approach is not directly scalable to voice disorders other than HD. The process of feature engineering is laborious and requires researchers with expertise in signal processing and machine learning as well as deep knowledge of the particular disorder and its underlying pathophysiological mechanisms. A model trained for one disorder will highly unlikely produce satisfactory predictions on data of another disorder. Even for the same disorder, the model's performance can differ greatly depending on the data acquisition conditions or labeling framework. For a more comprehensive list of factors affecting AVCA systems, refer to the work of Gómez-García et al. (2019) [16].

A DL approach might help to alleviate the problems with scalability and generalization, but and obstacle, as will be pointed out later, is a limited amount of available data, which is insufficient for today's DL models to fulfill their promises. The lack of data is a common denominator of almost all systems for automated medical data analysis.

> *"Deep Learning is getting really good on Big Data [...]. But Small Data is important too. [...] Hope more researchers work on Small Data – ML needs more innovations there."*
>
> *Andrew Ng, 2018*

Please, refer to the following articles for further information: *A Guide to Deep Learning in Healthcare (Esteva et al. 2019)* [13] provides a short, but comprehensive introduction to this topic, *A Survey on Deep Learning in Medical Image*

Analysis (Litjens et al. 2017) [31] provides an extensive survey regarding images, which is very relevant to this topic due to image-like properties of time-frequency representations of audio signals.

1.4 Objectives

From the perspective of the superordinate analysis of this dissertation, here I retrospectively delineate the main objective of this work, which is to **research new approaches to DL based predictive modeling using limited audio data sets, with a special focus on voice pathology assessment**. This main aim along with its sub aims will be later discussed in section Concluding Discussion. More specifically this dissertation aims to:

Aim 1: Explore the specifics of medical audio data analysis with DL
This constitutes conducting first experiments directly with the raw waveform in a search for an end to end system of voice pathology detection, which would map raw waveforms to the corresponding targets. Such experiments should also show the specific nature of the data and how to handle them with DL while determining the caveats.

Aim 2: Identify prospective DNN architectures w.r.t. AVCA systems
We plan to test popular DNN building blocks used in CV and in time-series analysis, namely CNN and LSTM, expecting automatic feature extraction.

Aim 3: Review available data sources and their limitations
More specifically to review their previous uses, identify which speech tasks they comprise, what is the distribution of healthy vs. dysphonic samples, what is the distribution of pathology types recorded and to propose an approach of combining the databases.

Aim 4: Clarify which input and target representations are useful
Specifically, to train models using raw waveforms and standard time-frequency representations, and compare the performances with handcrafted speech features. Moreover to identify, which other input modalities, such as gender, age, a grade of dysphonia, etc. affect the modeling capabilities and to suggest possibilities of redefining the task by changing the targets.

Aim 5: Propose countermeasures to high data demand
More precisely, to research and propose novel input and target data representations, which would benefit training on limited data sets.

2 Summary of the Publications

The main body of this work consists of five selected publications done during my doctoral studies. This section gives a short overview of their order, how the articles are contextually linked and how each of the preceding work and other events, like research visits, influenced the research direction and topic of the whole thesis. This timeline is presented in Table 2.1. The Publications are presented in versions of accepted or submitted manuscripts, their templates are unified, but contents are unchanged, apart from the numbering of tables, figures, equations and theorems, which may not fully reflect the official version.

Before I started to work on these articles, I did some prior work, where I was exploring the idea of DL and its application to time-series and audio data. In a paper entitled *Speech Emotion Recognition with Deep Learning (Harar, Burget & Duta, 2017)* [22], we have successfully used a CNN for automatic speech feature extraction and classification into one of three classes, i.e. emotional states – angry, neutral, sad.

After I was exposed to work and ideas of Mekyska and Galaz at the *Brain Diseases Analysis Laboratory*, I started to work on the utilization of DL in AVCA systems to avoid the "manual" feature engineering. Shortly after, I made a research visit to the University of Las Palmas de Gran Canaria, where Assoc. Prof. Jesús B. Alonso-Hernández generously provided his experience and further guidance.

Based on this cross-fertilization of ideas, a preliminary study entitled *Voice Pathology Detection using Deep Learning* [20] was published and presented at *International Conference and Workshop on Bioinspired Intelligence (IWOBI)* in July 2017. To the best of our knowledge, this was the first work in the world that studied the use of DL to solve this type of a problem. The objective of this study was to clarify, whether the use of DNN based on a combination of CNN and LSTM, applied to raw input audio signal, would prove itself worthy of further exploration for voice pathology detection. This work was chosen to be extended for a special issue in the journal *Neural Computing and Applications (IF 4.664, Q2 in AI)* and was once again presented at *Systematic Approaches to Deep Learning Methods for Audio* workshop in Vienna in September 2017.

The extended version with title *Towards Robust Voice Pathology Detection* [24] contains an extensive survey of previously published works, presents experiments conducted on four databases, namely Arabic Voice Pathology Database (AVPD) [37, 38], Massachusetts Eye and Ear Infirmary Voice Disorders Database (MEEI) [34], Príncipe de Asturias Database (PDA) [15] and SVD. Furthermore, it compares performances of ML and DL models trained using raw audio signal, spectral and cepstral time-frequency representations, and conventional handcrafted features. Also

Table 2.1: Timeline

- **Prior work**
  First published experiments with CNNs for audio sequence classification applied to speech emotion recognition.

- **University of Las Palmas de Gran Canaria (IDeTIC)**
  Acquired new data, exchanged ideas, and received guidance in the research of pathological voices from the machine learning perspective from Assoc. Prof. Jesús B. Alonso-Hernández.

- **Voice Pathology Detection Using Deep Learning**
- **Towards Robust Voice Pathology Detection**
  In-depth analysis of the state of the art and available data sets. Identified the main issues and conducted cross-database experiments.

- **University of Vienna (NuHAG)**
  Collaboration and supervision from Dr. Monika Dörfler in applied math and harmonic analysis. Strong focus on the fundamentals of neural networks and audio time-frequency representations.

- **On Orthogonal Projections for Dimension Reduction ...**
  Numerical experiments with augmented target loss function emphasizing important characteristics by beneficial representations of the target space.

- **Gabor Frames and Deep Scattering Networks in Audio ...**
- **Improving Machine Hearing on Limited Data Sets**
  Proposed and developed a software library for Gabor scattering and Mel scattering. Addressed the issue of insufficient amounts of data.

- **Future work**
  Combining the findings and applying them to voice pathology data.

Legend: – Journal article, – Conference paper, – Research visit

includes experiments with DenseNet [26] DNN architecture. It points out the limitations of the available data, the definition of the task and approach and suggests future work to alleviate the summarized problems.

In 2018, I have been awarded a grant for the mobility of researchers and thanks to the previously mentioned workshop in Vienna, I was given the opportunity to continue my research as a part of Numerical Harmonic Analysis Group (NuHAG) at the Faculty of Mathematics of the University of Vienna. This research visit under the supervision of Dr. Monika Dörfler radically changed my view on the problems at hand. I was invited to collaborate on multiple interesting fundamental research topics, for which I have conducted numerical experiments in music information retrieval setting, helped to design and implemented proposed algorithms, and created software libraries. In all the following articles, we have taken advantage of CNNs which were originally proposed for computer vision (CV), in predictive modeling with audio data. The reason is that standard FFT-based signal processing methods allowed exploiting advances in CV in the audio analysis by converting the raw audio waveforms into image-like representations (e.g. spectrograms).

A collaboration with the Department of Ophthalmology of the Medical University of Vienna led to an article accepted in the *Journal of Mathematical Imaging and Vision (IF 1.603, Q1 in CV)* titled *On Orthogonal Projections for Dimension Reduction and Applications in Augmented Target Loss Functions for Learning Problems* [7]. In this article, we studied the use of orthogonal projections on high-dimensional input and target data in learning frameworks and we introduced a general framework of augmented target loss functions (AT). These loss functions integrate additional information via transformations and projections of the target data. In two supervised learning problems, clinical image segmentation and music information classification, the application of our proposed AT increased the accuracy.

From the perspective of time-frequency analysis, in the paper *Gabor Frames and Deep Scattering Networks in Audio Processing* [3], we introduced Gabor scattering, a feature extractor based on Gabor frames and Mallat's scattering transform. Based on the provided theory, we have implemented the Gabor-scattering software library for Python programming language [19]. Furthermore, with numerical experiments, we showed, that the invariances encoded by the Gabor scattering transform lead to higher performance in comparison with just using Gabor transform, especially when few training samples are available.

As a next natural step, we included a human perceptual scale, which led to an extension of the Gabor scattering to a Mel scattering representation. The aforementioned software library was extended to cover both Gabor and Mel scattering. In the paper *Improving Machine Hearing on Limited Data Sets* [21] we investigated

how input and target representations interplay with the amount of training data in a music information retrieval setting. We compared the standard mel-spectrogram inputs with a newly proposed Mel scattering. Furthermore, we investigated the impact of additional target data representations by using the AT which incorporates unused available information. We observed that all proposed methods outperformed the standard mel-spectrogram representation when using a limited data set.

3 Concluding Discussion

To conclude this dissertation as a whole, the following section sums up the conclusions of the publications and is structured in such a way it tries to address the objectives in order of appearance in the section Objectives.

In the frame of **Aim 1** and **Aim 2**, we have hoped for an end to end system of voice pathology detection, which would map raw waveforms to the corresponding targets. The objective of the paper Voice Pathology Detection Using Deep Learning was to carry out a preliminary study which would clarify whether the use of the DNN model, especially combination of convolutional and LSTM layers would prove itself worthy of further exploration in case of voice pathology detection problem using only raw recordings of sustained vowel /a/. The examined method achieved 71.36 % accuracy on validation data and 68.08 % accuracy on testing data. It is important to note, that we did not restrict the classification to a subset of pathologies and we used all 71 present in the database.

We conclude that the main advantage of the DL approach with CNN is the automatic feature extraction, as opposed to the previously proposed methods. It saves a great amount of time and expertise in the area of the problem being solved. We found out, that the main disadvantage is the amount of data needed to train the model. The SVD database used in this experiment is extensive in numbers of persons recorded, but there are not enough samples of healthy persons in comparison with the number of samples of pathological patients. Also, the distribution of individual pathologies is extremely unequal making the voice pathology detection a hard problem.

In search of a robust voice pathology detection system using acoustic (voice) signals, researchers face a variety of problems. One of the major problems in this field of science, as we pointed out before, is the limited amount of data. Nevertheless, one large database from one source would not solve all the issues. A problem is also a limited number of distinct publicly available databases, using which the model could capture the variance of the data acquired in different recording conditions and environments. Following the **Aim 3**, the article Towards Robust Voice Pathology Detection explores publicly available data sources of dysphonic voices, discusses the means of combining them into one bigger database and uncovers their limitations concerning building an automatic assessment system.

The paper concludes, these commonly used databases (AVPD, MEEI, PDA, SVD) are very hard to combine because of various distinctions such as a) the databases are labeled in different languages, b) the databases do not comprise the same set of speech tasks, c) there is a variety of voice pathologies unequally distributed across the databases, etc. For these reasons, up to now, researchers have used only

a subset of the databases for their experiments providing results related to that carefully selected subset of data. However, this approach limits the possibilities of creating a robust voice pathology detector. We have conducted experiments on recordings of sustained phonation of the vowel /a/ produced at a normal pitch from the combination of these 4 different databases, trying to eliminate mentioned limitations. To the best of our knowledge, this is the first work that uses such a "large" set of data to build mathematical models for computerized, objective voice pathology detection.

To make a broader comparison, we researched 3 distinct classifiers within supervised learning and anomaly detection paradigms. Following the **Aim 4**, we have explored the usage of raw waveforms, spectrograms, MFCC, conventional dysphonic features and their combinations as input data. We observed that XGBoost classifier achieved the best results amongst DenseNet and Isolation Forest classifiers. In the article, we also investigated and described stratification and group weighting, to equalize the uneven distribution of gender-age groups, which is important to take into account, because of the different voice and speech properties of patients with different ages and gender.

Even though combining the available databases, we have obtained a relatively large amount of data samples, it still seems not to be enough to train a successful DL model on raw waveforms, and from the observed performances, we conclude that in voice pathology detection scenarios, with this (from AVCA perspective large, but from the DL perspective small) amount of training data, it is better to use inputs with reduced dimensionality in contrary to raw waveform inputs, and/or make use of transfer learning, data augmentation or other means to alleviate the problem with the lack of data. On the other hand, reviewing the performances achieved in scenarios with only MFCC as input data, we conclude that representations, as reduced in dimensionality as MFCC alone, are not reliable enough for robust voice pathology detection, which was also concluded by Ali et al. in [1].

We anticipate, that making the combination of the databases more controlled and coherent to reduce the noise in the database and simplifying the complexity of the target space would boost the performance of the system. Thus we think that recordings of the databases commonly used for automatic voice pathology detection should be consulted with clinicians as a whole, to evaluate the severity of vocal manifestation of the present pathologies based on perceptual evaluation as opposed to plain names of present pathologies. There are standard metrics, which are used to evaluate the quality of voice that can be used for this purpose [9, 10, 14, 28]. The addition of such information to the databases could provide researchers with a unique possibility to build models capable of classification and prediction, emphasizing the severity of the exact vocal-manifestation (increased acoustic tremor, roughness,

breathiness, etc.) of these pathologies.

At the end of the article, we anticipated that deep learning will play its role in robust voice pathology detection on the assumption that more data will be available, or at least reasonable combination of available databases will be made and limitations of these databases will be partially diminished by data augmentation and other countermeasures. Besides, we presume that the use of deep learning methods for novelty detection such as deep autoencoder [40] for modeling the normophonic voice could be an interesting idea for future investigation with a prospect to identify even disordered voices that are sparsely distributed across databases.

The first two publications were focused mainly on the specifics of predictive modeling using DL in voice pathology detection. They were concerned with identifying the prospective DNN architectures and dove deep into the analysis of available data sources. The following three publications all look at the problem of insufficient data, which was repeatedly mentioned in the first two publications, from a different perspective. As defined in the **Aim 5**, their objective is to propose methods of input and target space transformation in such a way, the DNN can learn with fewer data.

In the article On Orthogonal Projections for Dimension Reduction and Applications in Augmented Target Loss Functions for Learning Problems, we introduced a general framework of AT. These loss functions integrate additional information via transformations and projections of the target data. In two supervised learning problems, clinical image segmentation and music information classification, the application of our proposed AT increased the accuracy.

Next, in the article Gabor Frames and Deep Scattering Networks in Audio Processing, we introduced Gabor scattering (GS), a scattering transform based on Gabor frames and we investigated its properties. Thereby, we have been able to mathematically express the invariances introduced by GS within the first two layers. We have experimentally shown that explicit encoding of invariances by using an adequate feature extractor is beneficial when a restricted amount of data is available. It was shown that in the case of a limited data set the application of a GS representation improves the performance in classification tasks in comparison to using Gabor transform (GT). This property can be utilized in restricted settings, e.g. in embedded systems with limited resources or in medical applications, where sufficient data sets are often too expensive or impossible to gather, while the highest possible performance is crucial.

The common choice of a time-frequency representation of audio signals in predictive modeling is mel-spectrogram; hence, as a natural step, we introduced Mel scattering (MS) in Improving Machine Hearing on Limited Data Sets, a new feature extractor which combines the properties of GS with mel-filter averaging. We also investigated the impact of additional information about the target space through

AT on the performance of the trained CNN.

From the newly proposed methods, AT is the least expensive in terms of training time, but on the other hand, yields the smallest improvement in this experimental setup. Nevertheless, it has another advantage: it steers the training towards learning the penalized characteristics. We can conclude that AT provides a more precise measure of the distance between outputs and targets. That's why it can help in scenarios where the training set is not large enough to allow the learning of all characteristics but can be penalized by AT. We suggest using/experimenting with the proposed methods for other data sets if there is not a sufficient amount of data available or/and there exist reasonable transformations in the target space relevant to the task being solved. All proposed methods might be found useful also in scenarios with limited resources for training.

Beyond State of the Art This section concluding four long years of work is not short either, thus this paragraph briefly lists the achievements compactly:

- the first-ever use of deep learning in the field of voice pathology detection
- identification of limitations of deep learning w.r.t. this field
- identification of limitations of existing voice pathology databases
- experiments on the largest combined database of dysphonic voices
- design of new time-frequency representations based on Gabor transform
- improvement in the performance of convolutional neural networks on limited audio data sets using proposed novel time-frequency representations, namely Gabor scattering and Mel scattering, and a new class of loss functions, that yield beneficial target representations

Concurrent and Future Work The timeline in Table 2.1 constitutes only the main thread of my doctoral work, even though more work has been done during this period. Most notable are two collaborations: one with the Department of The Communication Disorders of the Comenius University in Bratislava. It is concerned with consulting available voice pathology databases combined into one, with clinical speech therapists, to evaluate the severity of vocal manifestation of the present pathologies based on perceptual evaluation according to GRBAS scale [9]. And the other, with the Austrian Research Institute for Artificial Intelligence (OFAI), experimenting with novel preprocessing steps for learning algorithms. During this collaboration, an experimental software library Redistributor [23] was developed. The results of these collaborations, unfortunately, did not make it into this work and are going to be worked upon and finalized in the future.

Bibliography

[1] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, A. Al-nasheri, T. A. Mesallam, M. Farahat, and K. H. Malki. Intra-and inter-database study for arabic, english, and german databases: Do conventional speech features detect voice pathology? *Journal of Voice*, 31(3):386–e1, 2017.

[2] E. Alpaydin. *Introduction to machine learning*. Adaptive Computation and Machine Learning series. MIT press, third edition edition, 2014.

[3] R. Bammer, M. Dörfler, and P. Harar. Gabor frames and deep scattering networks in audio processing. *arXiv preprint*, 2017. `arXiv:1706.08818`.

[4] A. Beavers. Alan turing: Mathematical mechanist. *Cooper, S. Barry; van Leeuwen, Jan. Alan Turing: His Work and Impact. Waltham: Elsevier*, pages 481–485, 2013.

[5] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2011.

[6] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova. Speech disorders in parkinson's disease: early diagnostics and effects of medication and brain stimulation. *Journal of neural transmission*, 124(3):303–334, 2017. `doi:10.1007/s00702-017-1676-0`.

[7] A. Breger, J. I. Orlando, P. Harar, M. Dörfler, S. Klimscha, C. Grechenig, B. S. Gerendas, U. Schmidt-Erfurth, and M. Ehler. On orthogonal projections for dimension reduction and applications in augmented target loss functions for learning problems. *Journal of Mathematical Imaging and Vision*, in press. `arXiv:1901.07598`.

[8] K. Cukier. Ready for robots: How to think about the future of ai. *Foreign Aff.*, 98:192, 2019.

[9] M. S. De Bodt, F. L. Wuyts, P. H. Van de Heyning, and C. Croux. Test-retest study of the grbas scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*, 11(1):74–80, 1997.

[10] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur Arch Otorhinolaryngol.*, 258(2):77–82, Feb. 2001.

[11] D. Doran, S. Schulz, and T. Besold. What does explainable ai really mean? a new conceptualization of perspectives. In *CEUR Workshop Proceedings*, volume 2071, 2018. URL: http://openaccess.city.ac.uk/id/eprint/18660/, arXiv:https://arxiv.org/abs/1710.00794.

[12] P. Enderby. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173, 1980. doi:10.3109/13682828009112541.

[13] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019. doi:10.1038/s41591-018-0316-z.

[14] B. R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, and G. S. Berke. Comparing internal and external standards in voice quality judgments. *J Speech Hear. Res.*, 36(1):14–20, Feb. 1993.

[15] J. I. Godino-Llorente, P. Gómez-Vilda, F. Cruz-Roldán, M. Blanco-Velasco, and R. Fraile. Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness. *Journal of Voice*, 24(6):667–677, 2010. doi:10.1016/j.jvoice.2009.04.003.

[16] J. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente. On the design of automatic voice condition analysis systems. part I: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51:181–199, 2019. doi:10.1016/j.bspc.2018.12.024.

[17] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[18] K. Gröchenig. Foundations of time-frequency analysis. 2001.

[19] P. Harar. Gabor scattering. https://gitlab.com/paloha/gabor-scattering, 2019.

[20] P. Harar, J. B. Alonso-Hernandez, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal. Voice pathology detection using deep learning: a preliminary study. In *2017 international conference and workshop on bioinspired intelligence (IWOBI)*, pages 1–4. IEEE, 2017. arXiv:1907.05905, doi:10.1109/IWOBI.2017.7985525.

[21] P. Harar, R. Bammer, A. Breger, M. Dörfler, and Z. Smekal. Improving machine hearing on limited data sets. In *2019 The 11th International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT)*. IEEE, in press. arXiv:1903.08950.

[22] P. Harar, R. Burget, and M. K. Dutta. Speech emotion recognition with deep learning. In *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 137–140. IEEE, 2017. doi:10.1109/SPIN.2017.8049931.

[23] P. Harar and D. Elbraechter. Redistributor. https://gitlab.com/paloha/redistributor, 2018.

[24] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal. Towards robust voice pathology detection. *Neural Computing and Applications*, pages 1–11, 2018. arXiv:1907.06129, doi:10.1007/s00521-018-3464-7.

[25] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. doi:10.1162/neco.1997.9.8.1735.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. URL: http://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html, arXiv:https://arxiv.org/abs/1608.06993.

[27] M. Košťálová, M. Mračková, R. Mareček, D. Beránková, I. Eliášová, E. Janoušová, J. Roubíčková, J. Bednařík, and I. Rektorová. Test 3F dysartrický profil–normativní hodnoty řeči v češtině. *Česká a Slovenská Neurologie a Neurochirurgie*, 76(109):5, 2013. URL: https://is.muni.cz/th/ucvby/8_Kostalova_3F.pdf.

[28] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear. Res.*, 36(1):21–40, Feb. 1993.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[30] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015. doi:10.1038/nature14539.

[31] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. doi:10.1016/j.media.2017.07.005.

[32] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017. doi:10.1016/j.neucom.2016.12.038.

[33] R. G. Lyons. *Understanding digital signal processing*. Pearson Education India, 2004.

[34] Massachusetts Eye and Ear Infirmary. Voice disorders database, version. 1.03 (cd-rom). *Lincoln Park, NJ: Kay Elemetrics Corporation*, 1994.

[35] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. doi:10.1007/BF02478259.

[36] Mekyska, J. Akustická analýza hypokinetické dysartrie u pacientů s Parkinsonovou nemocí: od základů až po integraci v mHealth systémech. XII. konference - neurogenní poruchy komunikace dospělých, Brno, 5 2017. FN Brno.

[37] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-nasheri, and G. Muhammad. Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. *Journal of healthcare engineering*, 2017. doi:10.1155/2017/8783751.

[38] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, A. Al-nasheri, and M. A. Bencherif. Voice pathology detection using interlaced derivative pattern on glottal source excitation. *Biomedical Signal Processing and Control*, 31:156–164, 2017. doi:10.1016/j.bspc.2016.08.002.

[39] A. V. Oppenheim and R. W. Schafer. *Discrete-time signal processing*. Pearson Education Limited, 2014.

[40] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[41] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):92, 2018. doi:10.1145/3234150.

[42] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019. doi:10.1109/JSTSP.2019.2908700.

[43] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. doi:10.1037/h0042519.

[44] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.

[45] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584. IEEE, 2015. doi:10.1109/ICASSP.2015.7178838.

[46] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. doi:10.1147/rd.33.0210.

[47] J. Schmidhuber. Deep learning. *Encyclopedia of Machine Learning and Data Mining*, pages 1–11, 2016. doi:10.1007/978-1-4899-7502-7_909-1.

[48] U. Zölzer. *Digital audio signal processing*, volume 9. Wiley Online Library, 2008.

Publications

I Voice Pathology Detection Using Deep Learning: a Preliminary Study

---

Bibliographic Information

P. Harar, J. B. Alonso-Hernandez, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal. Voice Pathology Detection Using Deep Learning: a Preliminary Study. In *2017 international conference and workshop on bioinspired intelligence (IWOBI)*, pages 1–4. IEEE, 2017. arXiv:1907.05905, doi:10.1109/IWOBI.2017.7985525.

Author's Contribution

The author surveyed related works, designed and performed the analysis, and wrote a significant part of the manuscript. He was also working on the finalization of the whole manuscript, i.e. reviewing, copyediting, etc.

Copyright Notice

Abstract

This paper describes a preliminary investigation of Voice Pathology Detection using Deep Neural Networks (DNN). We used voice recordings of sustained vowel /a/ produced at normal pitch from German corpus Saarbruecken Voice Database (SVD). This corpus contains voice recordings and electroglottograph signals of more than 2 000 speakers. The idea behind this experiment is the use of convolutional layers in combination with recurrent Long-Short-Term-Memory (LSTM) layers on raw audio signal. Each recording was split into 64 ms Hamming windowed segments with

30 ms overlap. Our trained model achieved 71.36 % accuracy with 65.04 % sensitivity and 77.67 % specificity on 206 validation files and 68.08 % accuracy with 66.75 % sensitivity and 77.89 % specificity on 874 testing files. This is a promising result in favor of this approach because it is comparable to similar previously published experiment that used different methodology. Further investigation is needed to achieve the state-of-the-art results.

Acknowledgment

This work was supported by the grant of the Czech Ministry of Health 16-30805A (Effects of non-invasive brain stimulation on hypokinetic dysarthria, micrographia, and brain plasticity in patients with Parkinsons disease) and the following projects: SIX (CZ.1.05/2.1.00/03.0072), and LOl401. For the research, infrastructure of the SIX Center was used.

II Towards Robust Voice Pathology Detection: Investigation of supervised deep learning, gradient boosting, and anomaly detection approaches across four databases

---

Bibliographic information

P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal. Towards robust voice pathology detection. *Neural Computing and Applications*, pages 1–11, 2018. arXiv:1907.06129, doi:10.1007/s00521-018-3464-7.

Author's contribution

The author significantly contributed to the survey of related works, obtained and prepared the data, designed and performed the analysis, and wrote a significant part of the manuscript. He contributed to each section of the article and organized the finalization of the whole manuscript, until its publication.

Copyright Notice

Abstract

Automatic objective non-invasive detection of pathological voice based on computerized analysis of acoustic signals can play an important role in early diagnosis, progression tracking and even effective treatment of pathological voices. In search towards such a robust voice pathology detection system we investigated 3 distinct classifiers within supervised learning and anomaly detection paradigms. We conducted a set of experiments using a variety of input data such as raw waveforms,

spectrograms, mel-frequency cepstral coefficients (MFCC) and conventional acoustic (dysphonic) features (AF). In comparison with previously published works, this article is the first to utilize combination of 4 different databases comprising normophonic and pathological recordings of sustained phonation of the vowel /a/ unrestricted to a subset of vocal pathologies. Furthermore, to our best knowledge, this article is the first to explore gradient boosted trees and deep learning for this application. The following best classification performances measured by F1 score on dedicated test set were achieved: XGBoost (0.733) using AF and MFCC, DenseNet (0.621) using MFCC, and Isolation Forest (0.610) using AF. Even though these results are of exploratory character, conducted experiments do show promising potential of gradient boosting and deep learning methods to robustly detect voice pathologies.

Acknowledgement

This study was funded by the grant of the Czech Ministry of Health 16-30805A (Effects of non-invasive brain stimulation on hypokinetic dysarthria, micrographia, and brain plasticity in patients with Parkinson's disease) and the following projects: SIX (CZ.1.05/2.1.00/03.0072), and LO1401. For the research, infrastructure of the SIX Center was used. The authors (P. Harar, Z. Galaz) of this study also acknowledge the financial support of Erwin Schrödinger International Institute for Mathematics and Physics during their stay at the "Systematic approaches to deep learning methods for audio" workshop held from September 11, 2017 to September 15, 2017 in Vienna.

III On Orthogonal Projections for Dimension Reduction and Applications in Augmented Target Loss Functions for Learning Problems

---

Bibliographic information

A. Breger, J. I. Orlando, P. Harar, M. Dörfler, S. Klimscha, C. Grechenig, B. S. Gerendas, U. Schmidt-Erfurth, and M. Ehler. On orthogonal projections for dimension reduction and applications in augmented target loss functions for learning problems. *Journal of Mathematical Imaging and Vision*, in press. `arXiv:1901.07598`

Author's contribution

The author prepared the data, designed and performed the analysis described in section Application to musical data, and wrote a report based on which the section was written.

Copyright Notice

Abstract

The use of orthogonal projections on high-dimensional input and target data in learning frameworks is studied. First, we investigate the relations between two standard objectives in dimension reduction, preservation of variance and of pairwise relative distances. Investigations of their asymptotic correlation as well as numerical experiments show that a projection does usually not satisfy both objectives at once. In a standard classification problem we determine projections on the input data that balance the objectives and compare subsequent results. Next, we extend

our application of orthogonal projections to deep learning tasks and introduce a general framework of augmented target loss functions. These loss functions integrate additional information via transformations and projections of the target data. In two supervised learning problems, clinical image segmentation and music information classification, the application of our proposed augmented target loss functions increase the accuracy.

Acknowledgement

This work was partially funded by the Vienna Science and Technology Fund (WWTF) through project VRG12-009, by WWTF AugUniWien/FA746A0249, by International Mobility of Researchers (CZ.02.2.69/0.0/0.0/16 027/0008371), and by project LO1401. For the research, infrastructure of the SIX Center was used.

IV Gabor Frames and Deep Scattering Networks in Audio Processing

---

Bibliographic information

R. Bammer, M. Dörfler, and P. Harar. Gabor frames and deep scattering networks in audio processing. *arXiv preprint*, 2017. `arXiv:1706.08818`, submitted.

Author's contribution

The author contributed to section Introduction, designed the implementation of the proposed algorithm and created most of the visualizations. Furthermore, he designed the synthetic data generator, preprocessed the data and performed the numerical experiments. Wrote a significant part of section Discussion and Future Work. He was helping with the finalization of the manuscript.

Abstract

This paper introduces Gabor scattering, a feature extractor based on Gabor frames and Mallat's scattering transform. By using a simple signal model for audio signals specific properties of Gabor scattering are studied. It is shown that for each layer, specific invariances to certain signal characteristics occur. Furthermore, deformation stability of the coefficient vector generated by the feature extractor is derived by using a decoupling technique which exploits the contractivity of general scattering networks. Deformations are introduced as changes in spectral shape and frequency modulation. The theoretical results are illustrated by numerical examples and experiments. Numerical evidence is given by evaluation on a synthetic and a "real" data set, that the invariances encoded by the Gabor scattering transform lead to higher performance in comparison with just using Gabor transform, especially when few training samples are available.

Acknowledgment

This work was supported by the Uni:docs Fellowship Programme for Doctoral Candidates in Vienna, by the Vienna Science and Technology Fund (WWTF) project SALSA (MA14-018), by the International Mobility of Researchers (CZ.02.2.69/0.0/0.0/16027/0008371), and by the project LO1401. Infrastructure of the SIX Center was used for computation.

V Improving Machine Hearing on Limited Data Sets

Bibliographic information

P. Harar, R. Bammer, A. Breger, M. Dörfler, and Z. Smekal. Improving machine hearing on limited data sets. In *2019 The 11th International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT)*. IEEE, in press. arXiv:1903.08950.

Author's contribution

The author preprocessed the data, designed and conducted the numerical experiments and prepared the visualizations. He wrote sections Numerical Experiments and Discussion and Conclusions and contributed to Introduction. Helped reviewing each section of the article and organized the finalization of the paper.

Copyright Notice

Abstract

Convolutional neural network (CNN) architectures have originated and revolutionized machine learning for images. In order to take advantage of CNNs in predictive modeling with audio data, standard FFT-based signal processing methods are often applied to convert the raw audio waveforms into an image-like representations (e.g. spectrograms). Even though conventional images and spectrograms differ in their feature properties, this kind of pre-processing reduces the amount of training data

necessary for successful training. In this contribution we investigate how input and target representations interplay with the amount of available training data in a music information retrieval setting. We compare the standard mel-spectrogram inputs with a newly proposed representation, called Mel scattering. Furthermore, we investigate the impact of additional target data representations by using an augmented target loss function which incorporates unused available information. We observe that all proposed methods outperform the standard mel-transform representation when using a limited data set and discuss their strengths and limitations. The source code for reproducibility of our experiments as well as intermediate results and model checkpoints are available in an online repository.

Acknowledgment

This work was supported by the Uni:docs Fellowship Programme for Doctoral Candidates in Vienna, by the Vienna Science and Technology Fund (WWTF) projects SALSA (MA14-018) and CHARMED (VRG12-009), by the International Mobility of Researchers (CZ.02.2.69/0.0/0.0/16 027/0008371), and by the project LO1401. Infrastructure of the SIX Center was used for computation.

Appendix

Curriculum Vitæ

Pavol Harár

PhD student @ Brno University of Technology
Researcher @ NuHAG, University of Vienna
Co-Founder & Researcher @ ACAI.AI

Main research interests

Analysis of pathological voices with deep learning
Exploration of novel time-frequency representations
Investigation of new preprocessing steps for learning algorithms

---

Personal

Residence	Vienna, Austria
Languages	Native in Slovak & Czech, English C1, French & German A1
Contact	pavol.harar@vut.cz, pavol.harar.eu

Education

2015 - *	PhD in Machine Learning, FEEC, BUT (expected in 09/2019)
2012 - 2015	Master of System Management and Informatics, FBM, BUT

Work experience

2019 - *	Researcher at NuHAG, University of Vienna
2019 - *	Co-Founder & Researcher at ACAI.AI
2015 - *	Researcher at Brain Disease Analysis Laboratory, BUT

Skills

Python, Numpy, Pandas, Keras, TensorFlow, Scikit-learn, Scipy, Matplotlib, Flask, Linux, Docker, Git, LaTeX, Java, Parallel & GPU computing

Teaching

2016 - 2017	Assistant lecturer: Basics of OOP in Java
2017	Coach: PyLadies Brno (public craschcourse of Python programming language initiated by PyLadies mentorship group)
2017	Bachelor thesis advisor: Vojtěch Hájek, Creating a database of audio recordings with artificial noise in an anechoic chamber
2016	Diploma thesis advisor: Martin Majtán, Trainable image segmentation usind deep neural networks

Participation in projects

2018 - 2019	Czech Ministry of Education, Youth And Sports (CZ.02.2.69/0.0/0.0/16_027/0008371): International mobility of researchers
2017 - 2019	Brno University of Technology (FEKT-S-17-4476) : Multimodal processing of unstructured data using machine learning and sophisticated methods of signal and image analysis
2016 - 2019	Ministry of Health of Czech Republic (NV16-30805A): Effects of non-invasive brain stimulation on hypokinetic dysarthria, micrographia, and brain plasticity in patients with Parkinson's disease
2015 - 2019	Vienna Science and Technology Fund (WWTF) (MA14-018): Semantic Annotation by Learned Structured and Adaptive Signal Representations (SALSA)
2015 - 2019	Czech Ministry of Education, Youth And Sports of Czech Republic (LO1401): Interdisciplinary research of wireless technologies (INWITE)

Internships

2018 - 2019	Numerical Harmonic Analysis Group (NuHAG), University of Vienna, Austria
2017	Technological Centre for Innovation in Communications (IDeTIC), University of Las Palmas de Gran Canaria, Spain

Awards

2017	The best lecturer at the Department of Telecommunications and 8^{th} best lecturer at the Faculty of Electrical Engineering and Telecommunications at Brno University of Technology

Invited lectures

06/06/2019	Hands on introduction to Attention mechanism at Deep Learning Seminar, University of Vienna, Austria
21/11/2018	Automatic Transformation of Empirical Data Distribution as an Experimental Pre-Processing Step for Neural Networks at Acoustic Research Institute (ARI), Vienna, Austria
25/06/2018	Basics of Neural Networks (Regression & Classification) at NuHAG, Vienna, Austria
05/06/2018	Towards Robust Voice Pathology Detection in a poster session of Harmonic Analysis and Applications Conference in Strobl, Austria
02/02/2018	Battle-proven machine learning workflow from venv to dockerization at BUT, Brno, Czech Republic
12/12/2017	How to train fox's brain to predict the future at Mergado DevTalks, Brno, Czech Republic
13/09/2019	Voice Pathology Detection Using Deep Learning: a Preliminary Study at Systematic approaches to deep learning methods for audio workshop, ESI, Vienna, Austria

Publications in journals

2019	Anna Breger, Jose Ignacio Orlando, Pavol Harar, Monika Dörfler, Sophie Klimscha, Christoph Grechenig, Bianca S. Gerendas, Ursula Schmidt-Erfurth, and Martin Ehler. On orthogonal projections for dimension reduction and applications in augmented target loss functions for learning problems. *Journal of Mathematical Imaging and Vision*, in press. arXiv:1901.07598.
2018	P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal. Towards robust voice pathology detection. *Neural Computing and Applications*, pages 1–11, 2018. arXiv:1907.06129, doi:10.1007/s00521-018-3464-7.
2018	Vojtěch Hájek, Pavol Hárar, Jiří Schimmel, and Radim Burget. But-czas: Korpus kvalitních nahrávek české řeči pořízených v bezodrazové komoře. *Elektrorevue*, 20(2):48–52, 2018.
2017	Roswitha Bammer, Monika Dörfler, and Pavol Harar. Gabor frames and deep scattering networks in audio processing. *arXiv preprint*, 2017. arXiv:1706.08818, submitted.

Publications in conference proceedings

2019 Pavol Harar, Roswitha Bammer, Anna Breger, Monika Dörfler, and Zdenek Smekal. Improving machine hearing on limited data sets. In *2019 The 11th Int. Cong. on Ultra Modern Telecom. and Control Systems (ICUMT)*. IEEE, in press. `arXiv:1903.08950`.

2018 Zoltan Galaz, Jiri Mekyska, Tomas Kiska, Vojtech Zvoncak, Jan Mucha, Pavol Harar, Zdenek Smekal, Ilona Eliasova, et al. Monitoring progress of parkinson's disease based on changes in phonation: a pilot study. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–5. IEEE, 2018. doi:10.1109/TSP.2018.8441307.

2017 P. Harar, J. B. Alonso-Hernandez, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal. Voice Pathology Detection Using Deep Learning: a Preliminary Study. In *2017 international conference and workshop on bioinspired intelligence (IWOBI)*, pages 1–4. IEEE, 2017. arXiv:1907.05905, doi:10.1109/IWOBI.2017.7985525.

2017 Pavol Harar, Radim Burget, and Malay Kishore Dutta. Speech emotion recognition with deep learning. In *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 137–140. IEEE, 2017. doi:10.1109/SPIN.2017.8049931.

2016 Garima Vyas, Malay Kishore Dutta, Jiri Prinosil, and Pavol Harar. An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pages 515–518. IEEE, 2016. doi:10.1109/TSP.2016.7760933.

Software

2019 Pavol Harar. Gabor scattering. https://gitlab.com/paloha/gabor-scattering, 2019.

2018 Pavol Harar and Dennis Elbraechter. Redistributor. https://gitlab.com/paloha/redistributor, 2018.

Scientific activity

H-index	3 (according to Scopus)
Citation count	20 (according to Scopus, excluding self-citations)