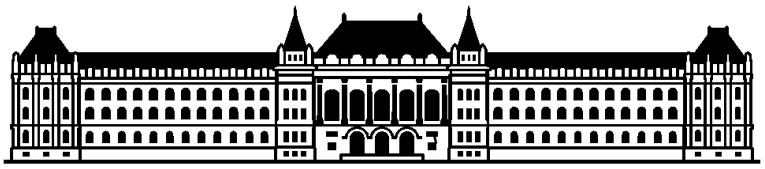

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
DEPARTMENT OF TELECOMMUNICATIONS AND MEDIA INFORMATICS

AUTOMATIC CLASSIFICATION OF DYSPHONIA

Ph.D. Thesis

Miklós Gábriel Tulics, MSc

Supervisor:

Klára Vicsi, DSc

BUDAPEST, HUNGARY

"Not everything is lost, as long as we are dissatisfied with ourselves."

— **Emil Cioran**

To whoever
didn't believe in me and
for those who did

Abstract

Dysphonia is a common complaint, reported in nearly one-third of the population at some point in their life and effecting almost every fourth child by producing a pathological voice. It affects the formation of clear and distinct sounds in speech as a complex function. Dysphonia is a pathological condition showing various symptoms due to several etiologic factors and pathogenesis diversity.

The knowledge of the acoustical features that describe pathological changes can help make a useful system for clinical practice. The system can provide clinical decision support whether a patient has dysphonia just by reading a short text. It would not only recognize dysphonic speech from a healthy one, but it would also recognize the type of dysphonia (namely: functional or organic dysphonia) and determine the severity of hoarseness. In this way the patient could be directed to the right health care facility faster and easier. The purpose of my research was to show that it is possible for adult's speech and for children's speech as well.

The structure of the dissertation is as follows: I briefly introduce the most important questions of speech processing regarding dysphonia, followed by my research objectives. I define dysphonia and its types, then characterize the type of change the illness causes in the speech product. I describe one of the most common scales used in medical practice to express the severity of dysphonia. I review the most common acoustic features, that can be suitable for classifying healthy and disordered speech and to predict the severity of the disordered voices automatically. I briefly list the artificial intelligence tools that are used in speech technology in this domain.

In the sections that follow I write about my own research. I describe the speech databases, statistical and machine learning methods I applied during my research. Based on the Hungarian Dysphonic and Healthy Adult Speech Database and statistical analyses, I confirm that the values of certain acoustic-phonetic features of speech change as a result of dysphonia. I describe the steps in the creation of a regression model that aims to determine the severity of hoarseness using the average rating which the four specialists used as target. Furthermore, I investigate the accuracy of the detection of dysphonia with different types of machine learning methods. I make a promising attempt on the automatic separation of functional and organic dysphonia, to my knowledge, there is no other research aiming to solve this difficult problem. I show that the methods and techniques used for the automatic classification of

dysphonic speech in adults can be applied to children's voice as well.

The obtained results in the theses allow us to implement a completely automatic diagnosis support system to recognize dysphonia in adults and children.

At the end of the study I give a summary about my results and theses.

Acknowledgements

First and foremost, I would like to thank my supervisor Klára Vicsi for many years of help, patience and perseverance. Thank you for believing in me, even when I did not believe in myself.

I am grateful to my family whose continuous support for all those years has always been well beyond the call of duty.

I would like to thank all the staff of the Laboratory Speech Acoustic group - especially Dávid Sztahó, Gábor Kiss, György Szaszák, Ildikó Nagy, - for their comments, criticisms and help in my work.

I would like to thank Krisztina Mészáros for her help in collecting the recordings over the last five years, for her professional advice and friendship. This research could not have taken place without her.

I would also like to thank Tamás Hacki, György Smehák and Nóra Damásdi for the RBH evaluation of the voice recordings. This was necessary for my research to create a model for the automatic estimation of the severity of dysphonia.

I would like to express my special thanks to Mária Ágostházy from the Speech Therapy and Vocational Education Service of Újbuda and Beke-Nádas Éva from the Cseresznyevirág Art Kindergarten for helping us construct the Dysphonic and Healthy Child Speech Database.

I am grateful to Krisztina Izabella Tulics, Hala Muhanna, Zsuzsanna Varga, Tamás Baltzer and Andrea Julianna Lajos for reading my Thesis, for the valuable suggestions and constructive criticism of the manuscript.

Last but not least, I would like to thank all the students I had contact with over the past few years for helping me grow as a person and as a researcher.

Key abbreviations

ANN ..... Artificial Neural Networks
ASR ..... Automatic Speech Recognition
DNN ..... Deep Neural Network
Dys ..... Dysphonia
ENT ..... Ear, Nose and Throat
EMD ..... Empirical Mode Decomposition
FD ..... Functional Dysphonia
FFS ..... Forward Feature Selection
HC ..... Healthy Control
HMM ..... Hidden Markov Model
HNR ..... Harmonics-to-Noise Ratio
ICC ..... Intra Class Correlation Coefficient
IMF ..... Intrinsic Mode Functions
LOOCV ..... Leave-One-Out Cross Validation
LPC ..... Linear Predictive Coding
MFCC ..... Mel-Frequency Cepstral Coefficients
OD ..... Organic Dysphonia
PER ..... Phone Error Rate
RBF ..... Radial Basis Function
ReLU ..... Rectified Linear Unit
RMSE ..... Root Mean Square Error
RP ..... Recurrent Paresis
SAMPA ..... Speech Assessment Methods Phonetic Alphabet
SPI ..... Soft Phonation Index
SVM ..... Support Vector Machine
SVR ..... Support Vector Regression

Contents

List of Tables

List of Figures

1 Introduction

Dysphonia (impaired voice production) generally refers to a condition where a person produces voice with an irregularity and it is affecting roughly 30% of the world's population at some point in their life [1, 2, 3, 4]. Dysphonia is not to be confused with hoarseness, as hoarseness is mostly reported by patients when they perceive an alteration in their voice quality, while dysphonia is recognized by a medical expert as hoarse, breathy, harsh or rough vocal qualities with a lower degree of phonation functionality.

Dysphonia affects patients of all ages, however research suggests that risks are higher in paediatric and elderly (>65 years of age) populations. Dysphonia is more common among professors, pedagogues, older adults and generally people who use their voice significantly more than the average in their professions [5, 6, 7, 8]. 23.4% of paediatric patients have dysphonia at some point during their childhood [9, 10, 11, 12]. The data therefore suggests that almost every fourth child produces a pathological voice. Studies agree that dysphonia is more often reported among boys than girls, the ratio being 70-30%.

People with various professions are affected by dysphonia but there is a higher likelihood of developing dysphonia among singers and entertainers, legal professionals, teachers, telemarketers etc [1, 13, 14, 15, 16, 17, 18]. Patients affected may experience an overall decrease in quality of life as it can affect a person's ability to work [19]. These people are in danger to miss work, lose wages, suffer from social isolation and develop depression. An incomprehensible speech limits a person's ability to communicate.

Dysphonia is classified as either an organic or a functional disorder of the larynx. Organic dysphonia (OD) results from some sort of physiological change in one of the subsystems of speech, while functional dysphonia (FD) refers to a voice problem in the absence of a physical condition. According to the American Speech-Language-Hearing Association, organic disorders can be subdivided into neurogenic and structural [20]. Neurogenic voice disorders include voice problems caused by abnormal control, coordination, or strength of voice box muscles due to underlying neurological diseases such as stroke, Parkinson's disease, multiple sclerosis, myasthenia gravis, and amyotrophic lateral sclerosis. Structural organic disorders include morphological alterations such as vocal cord nodules, polyps, gastroesophageal reflux disease (GERD), cyst and vocal cord paralysis (recurrent paresis, RP) [21].

2 Research objectives

I would like to contribute with my research to the speech-based detection of dysphonia and automatic estimation of its severity in the speech of adults and children by getting a deeper understanding of the effect of functional and organic dysphonia on speech. My specific goals during the research are:

a) Examining the possibilities of automatically detecting the severity of dysphonia;

b) Attempting a binary classification to separate dysphonic and healthy speech using different machine learning approaches;

c) Analysing the possibilities of automatically separating functional and organic dysphonia;

d) Analysing the possibilities of automatically separating healthy and dysphonic voices in children.

Contribution of my theses can be summarized as follows. I used Hungarian speech samples in all of my analyses. Hungarian is a language relatively poorly researched. In fact, the topic of automatic classification of dysphonic and healthy voices using Hungarian speech samples has not been studied yet, neither in adults, nor in children.

All my analyses are done in case of continuous speech. Sustained vowels might be easier to use because they do not require a resource intensive and language-dependent segmentation. However, they lack the information (such as prosody) that could be gathered from a continuous running (context rich) speech. Continuous speech has several advantages over the analysis of sustained vowels. For example, it contains variations of fundamental frequency, pauses and phonation onsets, and provides an opportunity to examine different variations of speech sounds. Treating continuous speech constitutes a new challenge, as it requires a different approach. However, due to its many advantages I adopted this paradigm in my research.

Furthermore, I tried to automatically separate functional and organic dysphonia. To my best knowledge, there has been no research aimed at the automatic separation of functional and organic dysphonia to this date. Having such a diagnosis supporting system that can separate not only healthy from dysphonic voices, but also can predict the type of dysphonia, can greatly accelerate the process in which patients are referred to specialists. If the system

detects functional dysphonia, the patient would be directed to a phoniatrist or speech therapist. However, if the system detects organic dysphonia, the patient would be directed to an otolaryngologists or oncologist. This could save a lot of time, leading the patient to care as soon as possible.

I tried to automatically separate the voices of healthy children from the ones with dysphonia. The end goal is to create a screening system that can be used by pre-school workers. If a child with dysphonic voice can be found on time, she or he has a better chance of getting professional help from an ear, nose and throat (ENT) specialist or a speech therapist.

3 Biomarkers, speech and dysphonia

3.1 Speech as an objective biomarker

Biomarkers are biological markers that indicate, among other things, the presence of a disease. They can be very diverse, such as inflammatory CRP (C-reactive protein) in the blood or a specific antibody that indicates the presence of an infection. Our voice is another biomarker which can refer to several diseases [22].

Speech is the most important way of human communication, but it contains much more information than the meaning of spoken words [23]. It tells a great deal about our mental and physical condition. If a person is sad, tired, has a cold etc. one might immediately hear it in his or her voice.

Speech production is a complex process that results in a mechanical wave propagating in a resilient medium that creates a sense of sound in the living beings.

When we speak, many of our organs are involved in the phonation process, so if any change occurs in these organs, then the properties of the produced speech will also change. If the change is audible and visible (for example in a waveform or in a spectrogram), it can be measured. This is possible by examining the acoustic features of the voice.

3.2 Speech production

Speech is an acoustic product with a linguistic meaning, but it also carries rich information that is not related to this linguistic meaning [23]. The physical signal corresponding to speech sounds is created by very precisely controlled physiological functions. Not only is speech generation a complex process, but the physical appearance of the signal produced, and the vibration of the speech itself demonstrates complexity and dynamical changes in frequency, time, and loudness.

The linguistic and non-linguistic characteristics of speech are simultaneously present in the complex acoustic signal generated during speech production. Linguistic features carry conceptual content expressed in written word forms, while non-linguistic features carry information such as the gender of the speaker, physical condition (fatigue, illness), or the emotional state of the speaker (joy, anger, amazement, excitement, sadness, disgust, etc.). These two components make speech feel more expressive and livelier than typed or written punctuation.

The basic physiological organs involved in speech production are divided into three areas. The lowest part is the subglottal space, where the lungs and trachea are located, followed by the glottal space with the larynx, and finally the supraglottal space where the pharynx, mouth and nasal cavity are located, also called the vocal tract. The functioning of each organ is controlled by the human brain. When the diaphragm contracts the lungs, air flows upward through the trachea to the larynx, which is the source of speech, the primary area of sound production. If we pronounce a voiced sound, the quasi-periodic opening of the vocal cords results in the formation of a quasi-periodic sound due to the opening and closing of the vocal cords. In the case of an unvoiced sound, the air passes unhindered through the larynx (blowing position). The air continues to pass through the oral and nasal cavities, where passive articulation organs (teeth, palate) and active articulation organs (tongue, uvula, lips, jaw) modify the waveform. Each individual phone has a specific articulation position of the articulation organs.

People's speech organs are not the same, they differ to a greater or lesser extent, which is why speech produced by humans is acoustically different (speech style, tone of voice). The acoustic structure of speech is influenced by the fact that articulation is not uniform across speakers. This is called the variation of speech between speakers (or sometimes inter-speaker variability). Furthermore, intra-speaker variability means that a person's acoustic product is not constant, beside a natural fluctuation it depends on the speech situation and on the current physical, health and emotional state of the person. Moreover, individual phones in continuous speech are connected by co-articulation during the phonation. Phones are connected to each other by a continuous transition. The transition parts are at least as important as those representing the phone‘s nuclei. During the co-articulation effect the pronunciation of phones is affected by the phones preceding and following them. This effect also causes intra-speaker variability of the acoustic characteristics of the phones.

The acoustic characteristics of the produced phones can be measured by objective methods. Due to the variations listed above, the acoustic characteristics of the individual's phones, and associated waveforms will not be the same, but in the best case they will be similar in healthy pronunciation. Human speech perception ensures that, despite these acoustic variation, phonological decisions can remain the same.

3.3 Dysphonia, and its relation to voice hoarseness

At a certain point in their lifetime almost one-third of the population is affected by a very frequent medical complaint that is called **dysphonia** [1, 2, 3, 4]. In common language it is called impaired voice production. The term dysphonia is often inter-charged with hoarseness, nevertheless this terminology is inaccurate because hoarseness is a symptom of altered voice quality reported by patients, while dysphonia is typified by a clinically recognized disturbance of voice production [24].

Its occurrence shows wide variety as it can affect patients of all ages and sexes, but shows an increased number among teachers, older adults and others with exceptional voice needs [5, 6, 7, 8].

Source of statistics say that annually 1 in 13 adults are somehow affected by voice problems [25]. A significant number of patients do not seek medical assistance even though the voice problems occurred and are recognized by the patients [25, 26, 27]. Dysphonia has a huge negative impact on productivity and economy, due to work absenteeism caused by the dense requisition of health care services [28]. Dysphonia is frequently caused by benign or self-limited conditions, but sometimes the existing symptom predicts a more severe or progressive condition, which requires immediate diagnosis and treatment.

A Clinical Practice Guideline was developed for hoarseness and dysphonia by Stachler and his colleagues [1]. The working definition developed by a guideline panel for dysphonia is the following: *"Altered vocal quality, pitch, loudness, or vocal effort that impairs communication as assessed by a clinician and/or affects quality of life"*, while hoarseness is defined as follows: *"A symptom of altered voice quality reported by patients"*. The panel's hypothesis is that this voice condition affects people diversely. The target population of this directive includes all individuals with dysphonia regardless of age. The guideline panel is extremely useful for all physicians who diagnose and treat patients with dysphonia. Henceforth the guide is applicable in any case where identifying, treating and monitoring dysphonia is targeted.

This guideline drew attention to many important things. Deficiency of knowledge and privity about dysphonia and its causes prevent the adequately care. For instance: ageing causes natural voice changes that are considered by older people as a natural process, while in this older age group some dysphonia may refer to symptoms of major and underlying diseases. Furthermore, parents misinterpret or consider hoarseness as being normal for their child. Such misinterpretations of symptoms hold back the proper evaluation, diagnosis and

treatment of some serious underlying health conditions. In order to minimize incorrect diagnoses and improve the quality of care, it is crucial to update the knowledge of health care professionals.

The guideline is useful for prevention by making recognition and diagnosis easier, as well as setting up targeted treatment for dysphonia. Furthermore, it points out and updates the needs and treatment options for specific populations. In a nutshell, the guideline's aim is to enrich the knowledge of professionals in the treatment, prevention and – through assessment – the promotion of appropriate treatment options for individuals affected by dysphonia.

3.3.1 Types of dysphonia

The concepts of **voice disorder** and **dysphonia** are not the same, yet many sources do not sharply separate the two. Voice disorder happens once somebody's voice quality, pitch, and loudness are inappropriate for an individual's age, gender, cultural background, or geographic location, while dysphonia bounds the auditory-perceptual symptoms of voice disorders [20]. Auditory-perceptual quality of voice in people with voice disorders will vary depending on the disease type and the severity of it. The severity of the voice disorder cannot always be determined based alone on auditory-perceptual voice quality [20, 29, 30].

According to the American Speech-Language-Hearing Association, voice disorders can be divided into two groups: organic voice disorders and functional voice disorders [20]. Organic disorders fall into two groups: structural and neurogenic [20]. Neurogenic voice disorders are caused by a problem in the nervous system, that include voice problems caused by abnormal control, coordination, or strength of voice box muscles due to an underlying neurological disease such as stroke, Parkinson's disease, multiple sclerosis, myasthenia gravis, and amyotrophic lateral sclerosis. Structural disorders on the other hand involve something physically wrong with the mechanism, such as alterations in vocal fold tissues such as oedema or vocal nodules, polyps, gastroesophageal reflux disease (GERD), cyst and vocal cord paralysis (recurrent paresis, RP) [21, 30].

From this we can state that dysphonia as well can be etiologically classified into two fundamental categories: functional dysphonia (FD) and organic dysphonia (OD) [20, 29, 30]. Organic dysphonias are the consequences of aspects non-related to the use of voice. Functional dysphonia refers to a voice problem without a physical condition resulting from poor vocal technique, and/or muscle imbalance, with or without psychoemotional involvement.

Organic dysphonia may occur from a large number of factors. For example, acute laryngitis can be caused by viral or bacterial infection, chronic laryngitis can be a result of heavy smoking or GERD. Trauma can be caused by surgery or intubation. On the other hand, function dysphonia is the result of improper or inefficient use of the vocal mechanism.

Symptoms of a patient suffering from dysphonia may present altered voice quality overall with hoarseness and a sore or dry throat. Other related symptoms may be for example loss of appetite and loss of weight, coughing up blood, acid reflux or heartburn, non-anginal chest pain and difficulty in swallowing [1, 30].

Any person who has a hoarse voice for four weeks or more is ought to seek for therapeutic thought from a doctor. Each condition has its own specific treatment, and the treatment needs to be customized to each individual by an ear, nose and throat (ENT) specialist or a speech therapist [1, 28, 30].

This traditional division of voice disorder classification is much criticized and is a rather vague division. Phoniatrist specialists and speech therapists do not deal with hoarse voices but with pathological voice formations. Therefore, the study of the formation of a "hoarse voice" is more of a physical-acoustic task [31]. Organic lesions also cause functional dysfunctions. Therefore, there are dysfunctions of organic origin and those of non-organic origin. Non-organic dysfunctions are criticized to call functional dysfunctions. The equivalent of inorganic dysphonia is not "functional" but "regulatory" [30].

A recent paper recommends the replacement of the term "functional dysphonia" with the term "malregulative dysphonia" [32]. Hacki and his colleagues argue that the term "functional" is by now so muddled and confused in everyday clinical usage and it is unclear what it means in any given clinical case.

3.3.2 Prevalence of dysphonia

Based on a 2001 US analysis the point incidence of dysphonia is 0.98% (536,943 patients with dysphonia per 55,000,000 patients). Corresponding to previous studies' proportion results, they noted higher rates among females (1.2% vs. 0.7% for males) and among those >70 years of age (2.5% vs. 0.6%-1.8% for all other age groups) [33, 34, 35, 36]. By the International Classification of Diseases, Ninth Revision, most frequently dysphonia related diagnosis made by medical professionals were acute laryngitis, non-specific dysphonia, benign vocal fold lesions (e.g., cysts, polyps, nodules), and chronic laryngitis. The real occurrence

of dysphonia related conditions are probably higher, because the majority of patients who suffer from voice changes do not seek treatment, especially if the symptoms are temporary and linked to an upper respiratory infection [33]. A previous study conducted a random survey in Utah and Iowa among adults who did not request medical treatment, and reported a 29.9% cumulative lifetime risk of a voice abnormality before 65 years of age.

3.3.3 Cost

Expenditures of dysphonia treatments are significant. According to a broad administrative database study direct cost of dysphonia treatments is estimated on average US $ 577 to US $ 953 per patient per year [37]. If the estimated 5.2 million affected patients by dysphonia would require treatment each year, the total direct health care would cost approx. US $13.5 billion. For comparison, these expenditures are similar to the amount spent on conditions such as chronic obstructive pulmonary disease, asthma, diabetes, and allergic rhinitis.

3.3.4 The impact on quality of life

Dysphonia primarily impairs the quality of life, unless it is the cause of a more severe condition (e.g., increased risk of death or morbidity). Impacts of dysphonia can affect the quality of life substantially. Limitations caused by dysphonia often bring side diseases as depression, social isolation, anxiety, missed work etc. [33, 38, 39]. Those studies which investigate voice abnormalities noted that quality of life consequences and work productivity losses are similar to those patients who suffer from other respiratory diseases like asthma, chronic obstructive pulmonary disease, acute coronary syndrome or even depression [26, 27]. In those cases when patients suffer from more serious types of the disease, e.g., unilateral vocal fold paralysis, a significant decrease in quality of life and productivity loss can be observed.

3.3.5 Dysphonia as symptom of an underlying disease

Dysphonia is a common symptom of many diseases. In case of head and neck cancer dysphonia may be an accompanied symptom, so it is crucial to recognise and pay attention to the early signs of it. Inaccurate examination of the larynx may delay the accurate diagnosis of cancer resulting higher stages, more invasive treatment and higher mortality rates [40]. Neurological conditions also bring forth dysphonia eg vocal fold paralysis (or recurrent paresis

- RP), spasmodic dysphonia (SD), essential tremor, Parkinson's disease, amyotrophic lateral sclerosis, multiple sclerosis, gastrointestinal (eg reflux, eosinophilic esophagitis), rheumatologic/autoimmune (eg rheumatic arthritis, Sjögren's syndrome, sarcoidosis, amyloidosis, granulomatosis with polyangitis), allergic, pulmonary, musculoskeletal (e.g., muscle tension dysphonia, fibromyalgia, cervicalgia), psychological (functional voice disorders), traumatic (e.g., laryngeal fracture, inhalational injury, iatrogenic injury, blunt/penetrating trauma), and infectious (e.g., candidiasis).

The incidence rate of dysphonia within these circumstances may of course change. For instance, almost all patients with SD or other laryngeal dystonia appear to have dysphonia, but not all patients who suffer from reflux have dysphonia.

3.3.6 Dysphonia and age

Voice changes and disorders affect all ages, but some proof points out there is a higher risk to occur among paediatric and ageing (>65 years of age) populations. According to estimated numbers, 23.4% of children have dysphonia at a certain point in their lifetime with increased incidence among boys and children between age 8 to 14 years [9, 10, 11, 12]. Incidence of dysphonia is also significantly higher between older adults with presbylarynx, which is an age dependent laryngeal mutation [34].

According to a huge broad administrative insurance claim database the percentage of elderly population requesting treatment for dysphonia was 1.3% (age range 60 to 69 years) and 2.5% among patients >70 years [2, 36]. In this group the most common diagnoses were acute chronic laryngitis, non-specific dysphonia, and laryngeal lesions. As stated by a previous survey made among elderly volunteers who forbore any kind of medical treatment, 47% suffered by a voice disorder during their lifetime and 29% were actively experiencing dysphonia [41]. Another study examined 120 elderly people at the Atlanta Volunteer Lifestyle Facility and found a 20% point prevalence of voice disturbance based on sound-related quality of life scores [42].

3.3.7 Dysphonia and profession

Among professions which require high vocal demands, dysphonia has a much higher chance to appear. In the most endangered position are singers and entertainers, legal professionals, teachers, coaches, clergy and anyone whose profession needs outstanding voice capacities

[1, 13, 14, 15, 16, 17, 18].

From this perspective it's axiomatic that voice diseases especially dysphonia have a big influence on a person's ability to work [19]. In the United States about 28 million workers live with voice problems on a daily basis [15]. From surveyed population in general 7.2% of the respondents missed work for ≥ 1 more days because of a voice problem, and 1 out of 10 fill short term disability claim [33]. In fact, 20% of teachers are absent from work due to dysphonia. This high rate of absenteeism just on this particular professional area causes $ 2.5 billion loss annually in the United States [15].

3.3.8 Environmental factors

In addition, environmental conditions can affect voice as well, such as background noise, poor air quality, and dry environments [43, 21, 44]. High levels of environmental or occupational irritations can also increase the likelihood of developing dysphonia, such as exposure to chemicals, smoke, dust particles, and contaminants.

One study examined 10 rescue workers working at the World Trade Center disaster site. The altered and hoarse voices of rescue workers were associated with the large amount of irritants present at the scene [45]. Dry or arid environments can also be harmful to our voice production.

At the same time, environmental humidification has had a beneficial effect on the dehydration of the superficial larynx, which may help prevent or reduce negative voice changes [46].

3.4 Inference

From the above subsections, it can be concluded that dysphonia affects a large number of people, with a third of the population affected by the disease during their lifetime. The change in voice affects not only the elderly but also the children. People who are more likely to use their voices during of their work (teachers, singers, etc.) are more prone to developing dysphonia, while the condition can affect a person's ability to work. Dysphonia can be caused not only by the frequent use of one's voice - it can be caused by many reasons, one of which is the constantly increasing environmental noise, environmental pollution, dry environment. The cost of treatment is a significant burden on health care.

Dysphonia can serve as an indicator to many serious and less serious diseases. For

example, hoarseness can be a sign of a life-threatening disease, such as laryngeal cancer. Moreover, hoarseness can also indicate vocal cord paralysis, that is often caused by lung cancer. Dysphonia also reduces people's quality of life in general and can lead to loneliness and depression. In conclusion, early diagnosis and therapy can be lifesaving.

The lack of knowledge and awareness about dysphonia and its causes are also potential blockers to receive appropriate health care. Older adults may perceive the changing of voice as part of ageing, but some symptoms may indicate a more serious underlying disease. Furthermore, a parent may misperceive hoarseness as being normal for his or her child. This assumption can slow down the exploration, diagnosis, and treatment of a child's condition. This is why improved education is needed on the topic among pedagogues and health professionals.

These facts all point to the need to provide physicians with tools to be able to diagnose dysphonia more easily and quickly at an early stage. It is advisable to separate the different types of dysphonia as soon as possible, thus speeding up the medical procedure. In addition, automatic determination of the severity of dysphonia would help to better map the patient's condition.

Educators who work with children would need tools to screen those with hoarseness. This way children could have access to a specialist or speech therapist as soon as possible. Such a system could even save lives. A voice-based diagnostic support system using continuous speech would all be suitable for this.

3.5 Scales of hoarseness and the RBH scale

Traditionally, the perception of voice quality of the listener has been attributed to a myriad of attributes that have developed under the influence of local linguistic traditions [30]. Associations often rely on impressions transmitted by other senses: dark, light, veiled, dull, sharp, and so on. Science therefore seeks to introduce voice categories that can be associated with the physiological-acoustic reality of phonation. However, advanced acoustic analysis is still unable to determine and classify the voice characteristics perceived by our hearing system in full detail, so we are (still) in need to rely on judgement of the ear. The most important area of auditory judgement, is hoarseness - which is the difference in voice quality from the usual, as stated in Section 3.3.

It is crucial for the judgement that the number of categories should not be large, but

rather clear. The differences between the categories should be substantial though. Above all, it is necessary for the experienced investigator to reproduce his or her judgements and for consensus to be established between several experienced investigators.

Hoarseness as a pathological category of voice quality is physiologically traced back to the *roughness*, as the irregularity of the vibration of the vocal cords and *breathiness* to the flow turbulences resulting from the inadequate closure of the vocal cords.

Hoarseness is measured by the GRBAS (Grade - overall judgement of hoarseness, Roughness, Breathiness, Asthenia, Strain) scale according to Hirano's study in 1981 in Anglo-Saxon and Japanese territory [47] [48]. Categories should be graded (G) from 0 to 3, where 0 is the lack of hoarseness (normal), 1 is a slight degree, 2 is a medium degree, and 3 is a high degree of hoarseness.

The powerlessness and tension categories describe the general tone of the patient. The Grade (G) gives the overall impression, an overall judgement of hoarseness, without any numerical addition. Based on clinical experience, the GRBAS scale is very useful and well applicable despite the fact that the distinction of the last two categories by hearing is sometimes difficult.

In the German-speaking area, the RBH scale has been widely used in auditing judgement of voice quality. The RBH scale gives the severity of dysphonia, where R stands for roughness, B for breathiness and H for overall hoarseness. The degree of the category H cannot be less than the highest rate of the other two categories. For example, if B = 3 and R = 2, H is 3, and cannot be 2 or 1. A healthy voice's code is R0B0H0; the maximum H and respectively RBH value is 3, so a voice's code with severe dysphonia is R3B3H3. Ptok and his colleagues demonstrated that the application of the RBH scale is suitable for clinical purposes [49].

The recorded voice samples in my experiments were classified by a leading phoniatrist according to the RBH scale [50]. This scale was used to differentiate the degree of voice disorders in the database. Speech examples of patients were labelled on the base of this numeric scale. In my work the overall hoarseness H was used. The interpretation of H is shown in Table 1. If H equals to 0 it can be said that the voice of the patient is not hoarse. If H equals to 1, it is interpreted as mild hoarseness, 2 as moderate, while 3 as severe hoarseness. A hoarse voice may indicate a serious underlying disease.

Another severity measure worth mentioning is the Dysphonia Severity Index (DSI) [51, 52]. This combines different features that describe voice, such as highest frequency and lowest intensity of the voice from the Voice Range Profile (often called the "phonetogram"),

Table 1: Interpretation of the H score.

Value of H	**Interpretation**
0	No hoarseness
1	Mild hoarseness
2	Moderate hoarseness
3	Severe hoarseness

jitter in percent and the maximum phonation time in seconds. Note that this measure can only be used with sustained vowels.

4 Related work - relationship between dysphonia and speech

4.1 Automatic classification of dysphonia

There are several approaches for the binary classification to separate samples of a healthy subject from voices affected by some disorder. The first question is whether to use sustained vowels or continuous speech. To date, researchers have achieved decent results in the separation of healthy speech and speech affected by some disorder using sustained vowels [53, 54, 55, 56, 57]. Sustained vowels have their advantages since they are easy to use because there is no need for a resource intensive and language-dependent segmentation. However, a significant proportion of researchers use continuous speech in their research [58, 59, 60]. Vicsi and her colleagues [58] point out that higher classification results can be achieved by using continuous speech. Continuous speech has many benefits over the analyses of sustained vowels providing fundamental frequency changes, pauses, phonation onsets, and it allows us to explore various differences of phones. The research findings are expected to be more applicable to practical work since continuous speech is used in real-world situations. Thus, a more natural way of voice production can be examined. For these reasons, continuous speech was used in form of read text material in my Thesis.

Several voice disorder speech databases exist for research use containing sustained vowels, such as the Arabic Voice Pathology Database (AVPD), the German Saarbrücken Voice Database (SVD) and the Massachusetts Eye and Ear Infirmary (MEEI) speech database. However, speech databases containing continuous speech is not easy to find. A carefully designed Hungarian dysphonic speech database that contains continuous speech is required.

It is worth noting, that Parkinson's disease can lead to a dysphonic voice (but not necessarily a hoarse one), so there are studies in which acoustic features are utilized in typical dysphonic speech research to recognize Parkinson's disease. Acoustic features like jitter, shimmer, Harmonics-to-noise ratio (HNR) (or Noise-to-harmonics ratio - NHR) values were taken into consideration. These acoustic features are also called "dysphonia measures" among other acoustic features [61].

Several types of input vectors are constructed for inspection, for example: x-vectors and i-vectors [62], glottal flow descriptors [63], auto-correlation and entropy features in different frequency regions [57], ASR posterior features [64] etc.

Choosing the most suitable classification algorithm poses another challenge. Deep neural networks usually require a lot of data, but there are examples in the literature [65, 66, 67, 59]. On smaller databases, however, we may want to try other classifiers that have good generalization capability on smaller data sets, such as SVM [58, 60, 57, 62].

Researches mainly focus on the binary classification between healthy and disordered speech [63, 65]. The classification results vary widely, depending on the type of dysphonia to be separated, the size of the speech database, the type of voice material used (sustained vowels or continuous speech), the pre-processing methods chosen and the classifier. Using sustained vowels usually lead to higher accuracies. In [55] researchers use the MEEI voice disorder database to test their designed and implemented health care software for the detection of voice disorders in non-periodic speech signals. The classification was done with SVM and the maximum obtained accuracy was 96.21%. In their experiment the sustained vowel /ah/ was used, vocalized by dysphonic patients and healthy people. In [57] accuracies are reported to be were 99.54%, 99.53%, and 96.02% for MEEI, SVD, and AVPD. In their study SVM was applied as a classifier and sustained vowel /a/ for both normal and pathological voices were extracted from the databases. Some of the works presented on the MEEI voice disorder database reveal very high results that led researchers to question the usefulness of the database. Muhammad et al. in [68] argues that the normal and pathological voice recordings are recorded in two different environments in this database. Therefore, it is hard to distinguish whether the system is classifying voice features or environments. Therefore, the results interpreted on the MEEI database should be treated with caution. Also, it points out the importance of the creation of an appropriate speech database to conduct such studies. Huiyi Wu and his colleagues [65] use sustained vowel /a/ for their examination of Saarbruecken Voice Database of 482 were healthy people and 482 patients who are diagnosed with organic dysphonia. A Convolutional Neural Network (CNN) was used for automatic feature extraction and classification between healthy and organic dysphonia. 88.5%, 66.2% and 77.0% accuracy were obtained on training, validation and testing data set respectively.

Using continuous speech, the task is more difficult, but these results are expected to be more applicable to real life practical work. In [59] researchers used the phrase “Guten Morgen, wie geht es Ihnen?” from the German Saarbrücken Voice Database to classify dysphonia and healthy voices. A 66% f1-score was reached in their experiment with Long-Short-Term-Memory and Convolutional Network for classification. The researchers point out that using Deep Learning to classify pathological speech requires a huge amount of data.

Vicsi and her colleagues in [58] used acoustic features measured from continuous speech to classify 59 speakers into healthy (26 speakers) and dysphonia (33 speakers) groups. Their classification accuracies were in the range of 86%-88%.

In some cases, researches focus on the separation of healthy speech and some specific disease, but some papers focus on multi class classification [59]. For example, in the work of Kazinczi et al. [69] researchers first show that an SVM based classifier can separate the healthy and pathological voices automatically with a relatively high 87% accuracy. Then a multi class classification was carried out providing 60% accuracy between healthy, FD and RP groups. Additionally, a two-class classification provided 85% accuracy between healthy and RP, 78% accuracy between FD and RP, 66% between healthy and FD groups. Another example is the work of Muhammad et al. [70] where researchers apply only formant values as features that aims the separation of multiple pathological diseases in a multi-class scenario. The setup aims to separate multiple organic dysphonia disease types, namely: cyst, GERD, paralysis, polyp and sulcus. Their classification rates are 67.86% for female and 52.50% for male speakers. Researchers of the referenced publication [59] not only performed binary classification between dysphonia and healthy speech but also attempted to classify dysphonia, laryngitis, paralysis of vocal cords and healthy voices at once. They obtained 40% of f1-score for the four classes, 67% between laryngitis and healthy and 80% between paralysis and healthy speech. In another interesting work [60] researchers created three classes: healthy subjects, subjects diagnosed with vocal fold oedema and nodules (physical disorders) and subjects diagnosed with unilateral vocal folds paralysis (neuromuscular disorder). Their main objective was to compare the performance of voice pathologies identification using continuous speech with the sustained vowel /a/ typically used in these applications. The results showed that continuous speech allows better results for the two systems and that Gaussian Mixture Model (GMM) classifier overcame the SVM classifier. GMM system reaches 74% accuracy rate while the SVM system obtains 72% accuracy. For the sustained vowel /a/, the accuracy achieved by the GMM and the SVM is 66% and 69% respectively.

To my best knowledge there is no study aiming at the automatic separation of FD and OD groups, although it would be a very important and interesting topic. On the other hand, separation of the two groups would speed up the screening of potentially more serious cases. In Thesis III I show that the separation of these two groups is feasible.

4.2 Automatic assessment of dysphonia

Objective assessment of vocal quality is a fundamental part of the evaluation process. Some subjective scales of dysphonia used in the medical field to quantify the severity of the disorder were discussed in 3.5. However, we face a difficulty in achieving reliable scores for severity, given that they depend on the internal standards of the examiner.

Specialists who deal with patients may have different sensitivity for specific dimensions (like roughness or breathiness), fatigue, attention, exposure to a range of voice pathologies and training in the perceptual qualification of vocal quality may also be factors that can lead to inconsistent judgements [71]. Due to these inaccuracies researches sought to correlate acoustic features with dysphonia severity scores. With today's technology, we have the opportunity to extract acoustic features from voice recordings and provide an objective estimate of the severity of dysphonia. This procedure provides non-invasive estimation and it is independent from the expertise of the medical professional.

Another question that arises is how to select the target variable for a regression procedure. Researchers perform regression based on the evaluation of one expert, or better, they take the mean of the evaluation of several professionals [72, 64, 73]. However, the consistency of professional evaluation has been examined only in a few cases. Due to the differences mentioned above, this can be a problem, therefore it is worth examining the evaluations of professionals in terms of consistency.

Furthermore, the already elicited problem of using continuous speech or sustained vowels has an impact on human rating. Studies revealed that listeners rate dysphonia, and especially breathiness, more severely in sustained vowels than in continuous speech [74, 75, 76, 77]. As mentioned before, using continuous speech is a more natural way of voice production.

In a research conducted by Maryn and his colleagues the relationship between acoustic features and the overall perceptual evaluations of vocal quality were examined [72]. Voice recordings from 22 vocally normal and 228 voice-disordered (both FD and OD) speakers were used in their experiment, both sustained vowel phonation and continuous speech. A model was developed using several acoustic features that aimed to determine the overall quality of voice based on the perceptual ratings of five experienced voice clinicians. Acoustic measures such as the smoothed cepstral peak prominence, HNR, shimmer local, shimmer local dB, slope of the long-term average spectrum, tilt of the trend line through the long-term average spectrum were used to build an equation (called the acoustic voice quality index) to quantify

the severity of overall dysphonia. The correlation of the mean ratings of the specialist of overall voice quality and the acoustic voice quality index resulted in 0.78. Although this research is promising, a fast clinical severity assessment can be troublesome if the patient needs to produce both sustained voice and continuous speech, as this prolongs the duration of the medical examination.

Further difference among researches is whether they determine the severity of dysphonia by classifying into severity groups [63, 64, 78] or by applying a regression procedure [52, 79, 80]. In [63] the AVPD database is used, by applying perceptual analysis the speech samples were graded into three severity levels: mild, medium and severe. In case of cyst the classification accuracy of mild, medium and severe severity categories was 73.8%, in case of paralysis 95.2%, in case of polys 67.5%. The experiment also shows that diseases can be classified according to severity groups with different efficiencies. The models trained may be highly dependent on the types of disease found in the training databases. In [64] researchers use utterance-level ASR posterior features as an SVM input to classify utterances into categories of mild, moderate and severe disorder. The two-class classification accuracy for mild and severe disorders is 90.3%, between mild and moderate disorders a significant confusion is observed. The goal of the research described in [78] was to distinguish grade, roughness, breathiness, asthenia, strain (GRBAS) scale scores of pathological voices directly by using a one-dimensional convolutional neural network (1D-CNN). They used an original dataset containing 1,377 voice samples of sustained phonation of the vowel /a/. Each voice sample was rated by three experts according to the GRBAS scale and the median values were used as the correct answer label. The accuracy and quadratic weighted Cohen's kappa was given as metric for the testing dataset. The accuracy for the G scale was 77.1% and substantial agreement (kappa equals to 0.710). The model for the R scale had an accuracy of 76% with moderate agreement (kappa equals to 0.536).

Since the RBH scale was developed on the basis of the GRBAS scale, we can find researches using both GRBAS [78, 52, 81] and RBH [80, 82] subjective scales. Both scales are widely applied by experts in the field and together they cover all types of voice disorders, regardless of their etiology [82].

4.3 The relationship between dysphonia and children's speech

Studies on the acoustic analyses of paediatric dysphonia are rare in literature, even though it is a very important field of interest. Approximately, 23.4% of children are affected by dysphonia [1]. Dysphonia has a negative impact on the physical and mental health of a child, as a dysphonic voiced child is unable to communicate properly, may face struggles in social and educational development and may also develop self-esteem and self-image issues [83]. Another major problem is that parents do not perceive dysphonia symptoms as vocal disorders in their children which can potentially delay diagnosis and proper medical care.

Since vocal nodules are the most common cause of dysphonia in children [84], the relationship of vocal nodules and changes of speech is in focus. A large study involving 254 patients documented a relationship between nodule size and severity of hoarseness, breathiness, straining, and aphonia [85]. Another study on 40 patients found no statistical relationship between nodules and objective or subjective voice measures. Objective measures included fundamental frequency abnormality, perturbation, shimmer, decreased respiratory support, air loss, or significant muscle tension [86].

Because the sample size of speech databases containing dysphonic children's voice is small, researchers focus mainly on statistical analyses rather than classification [87, 88, 89]. Correlation dimension (D2) values for sustained vowels were significantly higher in dysphonic children than in healthy children and analysis showed markedly higher percent jitter in dysphonic children than in healthy children [88]. Janete Coelho and his colleagues analysed the perceptual and acoustic vocal parameters of school age children with vocal nodules and compared them to a group without vocal nodules [90]. Five children were examined from both genders, aged from 7 to 12 years. The Mann–Whitney U test, with significance level $p < 0.05$ was used in their study. Statistically significant differences were registered between the group with vocal nodules vs. the group without vocal nodules, regarding the following parameters: fundamental frequency, shimmer, HNR, maximum phonation time for /a/ e /z/, s/z coefficient. The study of Gopi Kishore Pebbili and his colleagues [91] aimed to document the Dysphonia Severity Index (DSI) scores of 42 Indian children aged 8–12 years. DSI values were found to be significantly higher ($p=0.027$) in girls than in boys. DSI attempts to measure the severity of dysphonia based on the sustained production of a vowel, using a weighted combination of maximum phonation time, highest frequency, lowest intensity, and jitter of an individual. While using continuous speech t-tests revealed that

variations of jitter and shimmer, HNR and the first component (c_1) of the mel-frequency cepstral coefficients are good indicators to separate healthy and dysphonic voices in children [87].

More data would be beneficial for the development of a clinical diagnosis support system and for the assessment of paediatric voice disorders. To my best knowledge there is no study aiming the automatic separation of healthy and dysphonic voices of children. In Thesis IV I show that it is possible to separate these two groups with high accuracy using continuous speech.

5 Materials and methods

5.1 Databases

5.1.1 Dysphonic and Healthy Adult Speech Database

Voice samples from patients were collected during their appointments at the Department of Head and Neck Surgery of the National Institute of Oncology. Their speech was recorded in a quiet medical consulting room. Each patient was a native speaker of Hungarian, so there was no language barrier. All of them gave signed consent for their voices to be recorded.

It is known that the voice recording conditions affect the outcome of algorithms that measure the acoustic features used for classification or for the estimation of dysphonia severity. This influence is not to be underestimated. External or environmental noises, such as the fan noise of the computer, the air conditioner's noise, the noise of a cell phone ringing or vibrating, or sounds outside the recording room (eg traffic sounds or heavy winds) may all be captured by the microphone. For these reasons it is crucial to minimize noise in the recording room to attain an optimal signal-to-noise ratio (SNR). An SNR level of 30 dB is the minimum expectation to make a reliable recording worth analysing [92]. In order for the classification / regression algorithms to work well in a medical room, the voice recordings used for training must be also made in a medical room environment. While collecting the recordings this was considered as well.

The recordings were made using a near field microphone (Monacor ECM-100), Creative Soundblaster Audigy 2 NX outer USB sound card, with good quality A/D converter and low noise level (audio coding: PCM, sampling rate: 16 kHz, quantization: 16-bit). Each patient had to read out aloud one of Aesop's Fables, "The North Wind and the Sun". This folk tale is frequently used in phoniatrics as an illustration of spoken language. It has been translated into several languages, Hungarian included. The text is eight sentences long. The database was annotated and segmented on phone level with the help of an automatic phone segmentator which was developed in the Laboratory of Speech Acoustics [93] and was followed by manual corrections. The annotation was done using the SAMPA phonetic alphabet [94]. In the rest of this thesis, vowels and other speech sounds will be referred with SAMPA characters in brackets.

The collected speech database contains voices from people suffering from diseases such as tumours at various places of the vocal tract, gastroesophageal reflux disease, chronic

inflammation of larynx, bulbar paresis, amyotrophic lateral sclerosis, leucoplakia, spasmodic dysphonia, etc. The most frequent diseases are functional dysphonia (referred to as 'FD') and recurrent paresis (referred to as 'RP'). We refer to the recordings from patients with dysphonia as 'Dys'. Recordings from healthy control were collected as well. These recordings are used as comparison, and they were collected from people who had attended for unrelated check-ups. We refer to these recordings as 'HC'.

The distribution of the voice recordings in the database is shown in 2. The database contains a total of 450 recordings, 257 from patients with dysphonia (156 females and 101 males) and 193 from people with a healthy voice (108 females and 85 males).

In the course of my research the database was constantly expanding with new recordings, this is why I drew my conclusions from a smaller database in case of some thesis statements. At each thesis point I present the database I used.

Table 2: Dysphonic and Healthy Adult Speech Database.

	Diagnosis		
Sex	Dysphonia	Healthy	Total
Female	156	108	264
Male	101	85	186
Total	257	193	450

5.1.2 Dysphonic and Healthy Child Speech Database

Voice samples from children were collected at several kindergartens. All the recordings were made with parental consent, mostly in the presence of the children's parents. The children recited a poem entitled "The Squirrel", written by Erika Bartos. This poem was chosen for therapeutic reasons, speech therapists use the it during treatment. As children in the 5-10 year old age group are very fond of the poem, it is easy for them to learn it. The most frequent vowel in the poem is the vowel [o], with 16 pieces followed by 14 pieces of the vowel [O] and 9 pieces of vowel [E].

The recordings were created using a near field microphone (Monacor ECM-100), Creative Soundblaster Audigy 2 NX outer USB sound card, with 44.100 Hz sampling rate and 16-bit linear coding. The duration of the recordings is about 20 seconds each.

The segmentation was made with the help of an automatic phone segmentator (mentioned in Section 5.1.1), followed by manual corrections. A total of 59 recordings were used in this

work: 25 voices from children with dysphonia (mean age: 6.52(±1.94)) (3 children had vocal nodes, the rest had functional dysphonia) and 34 recordings from healthy children (mean age: 5.35(±0.54)). Table 3 summarizes the recordings from the database used in the experiments.

Table 3: Dysphonic and Healthy Child Speech Database.

	Diagnosis		
Sex	Dysphonia	Healthy	Total
Girl	5	15	20
Boy	20	19	39
Total	25	34	59

5.2 Creating the input vector

5.2.1 Input vector from acoustic features

This section provides a description of the acoustic features I applied in order to process speech signals to extract clinically useful information. These features are widely used in the speech signal processing research community and have been developed for various purposes, including speaker identification, speech recognition, speech synthesis etc., and -as is the case with the present study- for extracting clinically relevant information. Other tools have been developed within distinct but related disciplines such as time series analysis.

When I examined adults' voice, I created the input vector for the classifiers used from acoustic features measured on vowel [E] (being the most frequent vowel in the read text), on different phonetic classes and on the whole wave file.

For vowel [E] jitter(ddp), shimmer(ddp), HNR (Harmonics-to-Noise Ratio) and the first component (c_1) of the mel-frequency cepstral coefficients (referred to as 'mfcc01') were measured. **On different phonetic classes** Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based IMF entropy frequency band ratios were measured on the voiced parts of speech, and the measured features were grouped into different phonetic classes: the vowel [E], nasal phones marked with [m], [n] and [J], on high vowels marked with [E], [e:], [i], [2] and [y], on low vowels marked with [O], [A:], [o] and [u], voiced spirants marked with [v], [z] and [Z], voiced plosives and affricates marked with [b], [d], [g], [dz], [dZ] and [d']. SPI was calculated on the **whole sample** as well. **Derived (utterance level) acoustic features** were calculated as the means, standard deviations and ranges of acoustic features (range is defined as max - min of raw values). In this way, a total of 49 acoustic

features were measured per each patient's voice sample, so 49 dimensional input vector was prepared from acoustic features. This feature set is further referred to as 'the 49 feature set'.

When dealing with children's speech, acoustic features were measured on vowel [o] (being the most frequent vowel in the poem), on different phonetic classes and on the whole wave file.

For vowel [o] the following acoustic features were measured: jitter, shimmer, HNR (Harmonics-to-Noise Ratio), 12 mfccs, the fundamental frequency (F0), formant frequency (F1, F2, F3), Formant frequency bandwidth (F1BW, F2BW, F3BW). **On different phonetic classes** SPI and IMF entropy frequency band ratios were measured on the voiced parts of speech, and the measured features were grouped into different phonetic classes: on vowel [o], on nasal phones, on high vowels, on low vowels, on voiced spirants, on voiced plosives and affricates. SPI was calculated on the **whole sample** as well. **Derived acoustic features** were calculated as the means, standard deviations and ranges of acoustic features. A total of 103 acoustic features were calculated per each children's voice sample.

A brief description of the used acoustic features follows.

Jitter [%]

Jitter (ddp) is the average absolute difference between the duration of consecutive time periods (T) in speech, divided by the average time period. The abbreviation 'ddp' refers to Difference of Differences of Periods. Calculation of jitter goes as follows:

$$Jitter_{ddp} = \frac{\sum_{i=2}^{N-1} |2 \cdot T_i - T_{i-1} - T_{i+1}|}{\sum_{i=2}^{N-1} T_i} \cdot 100[\%]. \quad (1)$$

T_i is the duration of the i-th interval and N is the number of intervals.

Shimmer [%]

Shimmer (dda) is the mean absolute difference between consecutive amplitude differences of consecutive periods. The abbreviation 'dda' refers to Difference of Differences of Amplitudes. Its calculation goes in a similar way:

$$Shimmer_{dda} = \frac{\sum_{i=2}^{N-1} |2 \cdot A_i - A_{i-1} - A_{i+1}|}{\sum_{i=2}^{N-1} A_i} \cdot 100[\%]. \quad (2)$$

In the formula, A_i is the magnitude of the i-th period and N is the number of intervals.

Harmonics-to-Noise Ratio (HNR) [dB]

HNR represents the degree of acoustic periodicity. It is calculated with the following formula

$$HNR = 10 \cdot \log \frac{E_H}{E_N}[dB], \tag{3}$$

where E_H is the energy of the harmonic component, while E_N is the energy of the noise component.

When measuring jitter, shimmer and HNR a pitch calculation time step of 0.01 seconds and calculation window of 0.064 seconds was used.

Mel Frequency Cepstral Coefficients (MFCCs)

MFCCs are widely used in automatic speech and speaker recognition, where frequency bands are equally spaced on the mel scale, that approximates the human auditory system's response. To calculate the MFCCs on needs to do the following steps:

1. First, we need to frame the speech signal into short frames (usually into 20-40 ms frames, 25 ms is standard). Frame step is usually 10 ms, which allows some overlap to the frames.

2. For each frame we need to calculate the periodogram estimate of the power spectrum.

3. Then we apply the mel filter bank to the power spectra. The Mel-spaced filter bank is a set of 20-40 (26 is standard) triangular filters that we apply to the periodogram power spectral estimate.

4. We sum the energy in each filter and take the logarithm of all filter bank energies (P_j).

5. MFCCs are the output of a Discrete Cosine Transform (DCT) on spectral values P_j. Usually we keep the 1-12 DCT coefficients and discard the rest. DCT is given by the following equation:

$$c_{k-1} = \sum_{j=1}^{N} P_j \cdot \cos(\frac{\pi \cdot (k-1)}{N} \cdot (j - 0.5)), \tag{4}$$

where N represents the number of spectral values and P_j the power in dB of the j-th spectral value (k runs from 1 to N). In case of adults the first component (c_1) of the mel-frequency cepstral coefficients (referred to as 'mfcc01') was used. When measuring MFCCs the time step was 0.01 seconds and a window size was 0.025 seconds.

Soft Phonation Index (SPI)

Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based IMF entropy frequency band ratios were measured on the voiced parts of speech, and the measured features were grouped into different phonetic classes: nasal phones marked with [m], [n] and [J], on high vowels marked with [E], [e:], [i], [2] and [y], on low vowels marked with [O], [A:], [o] and [u], voiced spirants marked with [v], [z] and [Z], voiced plosives and affricates marked with [b], [d], [g], [dz], [dZ] and [d']. SPI was calculated on the whole sample as well; no standard deviation was calculated here.

SPI is the average ratio of energy of the speech signal in the low frequency band (70-1600 Hz) to the high frequency band (1600-4500 Hz).

$$SPI = \frac{E_{70-1600Hz}}{E_{1600-4500Hz}}[dB]. \tag{5}$$

A time step of 0.01 seconds and a window size of 0.025 seconds was used during calculations.

Empirical mode decomposition (EMD) based IMF entropy frequency band ratios

Empirical mode decomposition (EMD) was proposed by Huang et al. for analysing non-stationary and non-linear time series [95] [96]. EMD can be combined with Hilbert spectral analysis that is called Hilbert - Huang Transform (HHT). HHT preserves the characteristics of the varying frequency which is an important advantage in real-world signals. In the EMD process a multicomponent signal is decomposed into elementary signal components called intrinsic mode functions (IMFs). The decomposition is based on the local characteristic time scale of the data, it is applicable to nonlinear and non-stationary processes, such as biomedical signals like speech. IMFs represent zero-mean amplitude and frequency modulated components [97]. EMD is a fully data-driven, unsupervised signal decomposition. EMD is similar to Fourier and wavelet analysis as it also satisfies the perfect reconstruction property, i.e. the extracted IMFs together with the residual slow trend reconstructs the original signal

without information loss or distortion. The most challenging problem of the procedure is to interpret the extracted IMFs in physical terms. During the algorithm, the IMFs are arranged in a matrix in sorted order according to frequency.

The EMD algorithm can be summarized in five steps:

1. Define the minimum and maximum of the signal (in our case waveform);
2. Use cubic spline interpolation to create a bottom envelope from connected minimums and an upper envelope from connected maxima;
3. Calculate the mean time series of the envelopes;
4. Subtract the mean time series from the data to obtain the next IMF component;
5. Repeat steps 1-4 so that the starting signal is the signal from the previous step.

Stop the process when the received signal meets the given end criterion.

An example of the EMD decomposition of a non-stationary time series $S(t) = cos(7t) + sin(4t) + 0.1t$ can be seen in Figure 1, while a decomposition of a [E] phone from continuous speech in Figure 2.

In Figure 1 the signal constructed by a sinus and cosine component was generated for 10 seconds, the EMD decomposition provided four IMFs. While the signal showed in Figure 2 is from the first [E] phone from the Hungarian word “egyszer”. The length of the phone is 70 ms, the EMD decomposition provided seven IMFs.

The first few IMFs represent the high frequency components of the signal, the latter IMFs represent the lower frequency components. I calculate the entropy (H) for each IMF. The frequency band ratios of entropy were calculated the following way:

$$IMF_{entropy} = \frac{\sum_{d=1}^{2} H_d}{\sum_{d=2}^{D} H_d}. \tag{6}$$

H_d is the value of Shannon entropy for each d = 1,2,...D of the log-transformed IMFs. D is the total number of extracted IMFs.

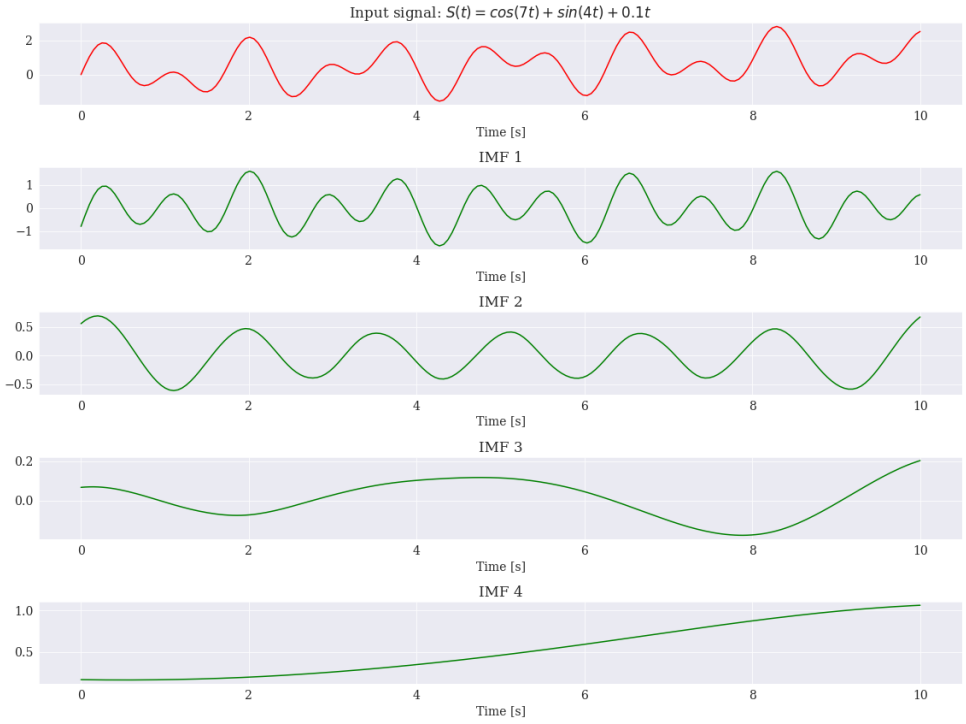

Figure 1: An EMD decomposition of the non-stationary time series $S(t) = cos(7t) + sin(4t) + 0.1t$.

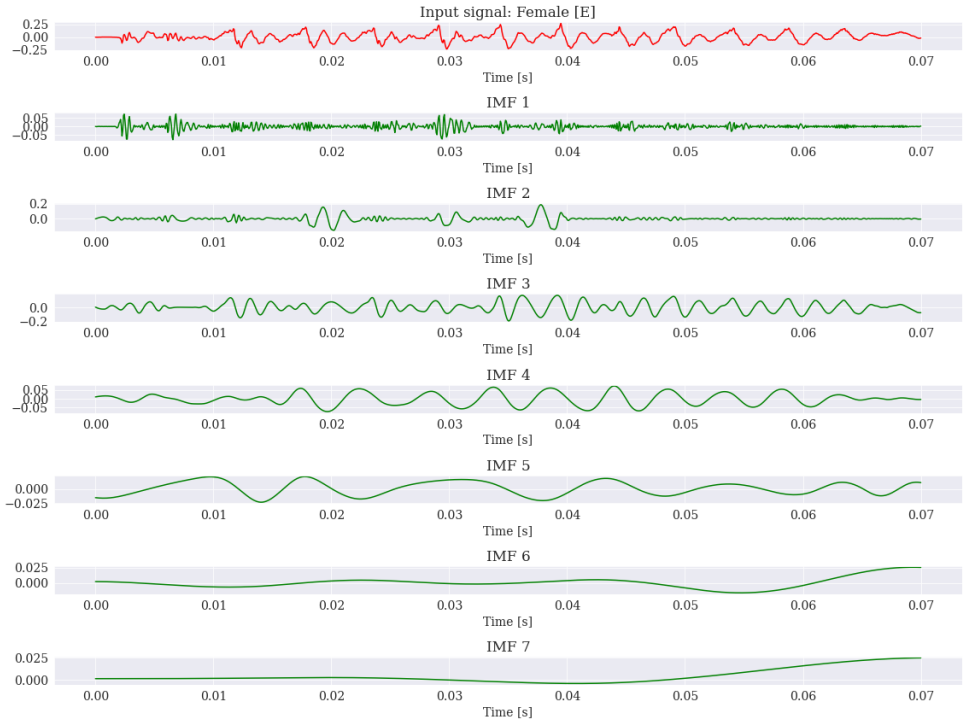

Figure 2: An EMD decomposition of a female [E] phone from continuous speech.

For a discrete random variable x, which can take a finite number n of possible values $x_i \in \{x_1, \ldots, x_n\}$, with probabilities of $p_i \in \{p_1, \ldots, p_n\}$, the Shannon entropy is defined as

$$H(x) = -K \sum_{i=1}^{n} p_i log p_i, \tag{7}$$

where K is a positive constant.

The entropy of a signal is calculated from the distribution (histogram) of the amplitude values, which can be considered as a density function. The value of p_i can be interpreted as the 'probability of occurrence' for a given amplitude.

In this way, a total of 49 acoustic features were measured per each patient's voice sample, so 49 dimensional input vector was prepared from acoustic features.

Fundamental frequency (F0) [Hz]

Human voices can be divided into two groups: voiced and unvoiced. In case of voiced sounds when the vocal cords open and close quasi-periodically as a function of the change in pressure conditions, the resulting amplitude-time function can be regarded as periodic or quasi-periodic. The time taken for the vocal folds to complete one oscillation (cycle) is known as pitch period T (sec). The fundamental frequency is the reciprocal of this period, thus its value is given by the following formula:

$$F0 = \frac{1}{T}[Hz]. \tag{8}$$

Speakers have a wide range of fundamental frequencies, but the average range of fundamental frequencies is different for children (250-500 Hz) [98], women (150-300 Hz), and men (80-150 Hz) [99] [23]. This is because men and women have different vocal fold sizes, which causes different patterns of vibration, in particular the number of times the vocal folds vibrate during a second. This is also true for children as their vocal organs are not fully developed yet.

The fundamental frequency calculation was done by an autocorrelation method described in [100] with a pitch calculation time step of 0.01 seconds and a calculation window of 0.064 seconds.

The first, the second and the third formant frequency (F1, F2, F3) [Hz]

In the case of voiced sounds, the vibration of the vocal cords excites the articulation channel. The sound produced at the vocal cords can be resolved to the fundamental frequency and its harmonics (integer multiples of the fundamental frequency), up to about 5000 Hz in healthy cases. In general, a decrease of 12 dB / octave in the amplitude values of the harmonic structure is observed. The fundamental frequency has the highest amplitude. This harmonic structure enters the articulation channel, where amplifications can be observed in certain areas of the spectrum, around the resonant frequencies generated by the articulation channel. Such amplified spectral areas in the spectrum are called formant frequencies. The frequency of a given formant refers to the location of the maximum amplitude of the envelope of the resonance frequency range [23].

Each voiced sound has its typical formant frequency values. I case of vowels the first and second formant frequencies clearly define the vowel. If, for a given sound, the values of the formant frequencies differ significantly from the formant values of the sound specified by the language, it is considered an articulation error.

Formant frequency tracking was realized by applying Gaussian window for a 150 ms long signal at a 10 ms rate. For each frame LPC coefficients were measured.

The bandwidths of the first, the second and the third formant frequency (F1BW, F2BW, F3BW) [Hz]

Bandwidth in signal processing is defined as the frequency region in which the amplification differs less than 3 dB from the amplification at the centre frequency. This is the range of a lower and upper frequency cut-off values. In speech technology the term is the same as in signal processing, formant frequencies being the centre frequencies.

The bandwidths of each formant frequency are denoted as F1BW, F2BW, F3BW and so on, similarly to formant frequencies. The higher the formant number, the broader the bandwidth is generally [23]. Increased bandwidth values compared to normal speech indicate that the amplification of formant frequencies is inadequate, which is an articulation error.

5.2.2 Input vector from phone level posterior probability values of an ASR

The acoustic models of Automatic Speech Recognizers (ASR) can also be used to extract features for dysphonia detection and classification. Today‘s state-of-the-art hybrid ASR acoustic models are composed of a transition model (a Hidden Markov Model) and a phone

classifier (DNN) [101]. The phone classifier can also be used to classify frames in standalone mode (without adding the recognition network and the ASR decoder) by using a forward pass for the speech frames one-by-one. In this way we obtain posterior probabilities of phones every 10 ms time frame from the DNN softmax layer of the phone classifier. Hence, only the phone classifier component of the acoustic model is used for prediction.

The acoustic model used for my experiments was trained on Hungarian data mixed from BABEL [102], the Hungarian Reference Speech Database (MRBA) [103] and the Hungarian Broadcast News Database [104] with the Kaldi toolkit [105], following the 'nnet2' WSJ recipe. Its phone classifier is based on spliced and LDA+MLLT transformed MFCC features input into a feed-forward DNN with 4, 1024 dimensional hidden layers with p-norm nonlinearity (p=2) and a softmax output for up to 2500 senones (context sensitive logical phone entities). After the forward pass in inference time, the senones are collapsed to phones of the 39 element SAMPA Hungarian phone set [102].

When the phoneme in question takes the highest probability value of all the other phonemes in the frame (i.e., the phoneme 'wins its frame'), I stored the values in a list, then I calculated the mean, standard deviation and the range. I did the calculations for vowels [E], for nasals, high vowels, low vowels, voiced spirants and plosives and affricates, for each recording. This resulted in a 21 dimensional vector per recording. I refer to this input vector as "ASR posterior features" to maintain coherence of terms with international literature, although as we have seen, these features are not ASR features in the strict sense, as they are generated by the phone classifier of a small acoustic model.

5.3 Statistical methods

5.3.1 Chi-squared tests

The chi-square test is a non-parametric statistical technique [106]. The chi-square test uses frequencies of categorical or ordinal data. The test can be used to determine the relationship between categorical variables (test of Independence). Test statistics that follow a chi-square distribution are independent and normally distributed. This assumption is often justified due to the central limit theorem. The chi-square assumes that the data for the study is a random sample from a population and the categories are mutually exclusive.

The chi-squared statistic summarizes the differences between the expected frequency when an outcome occurs and the observed frequency of it. This is done by summing the

squares of the differences, then normalized by the expected values over the categories. This can be calculated with equation 9.

$$\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}, \tag{9}$$

where O is the observed values (data), E is the expected values, k is the index and n is the number of mutually exclusive classes. The null hypothesis ($H0$) for a chi-square test is that the observed values and the expected values are independent, the alternative hypothesis ($H1$) states that they are dependent. The test calculates a p-value. If p is greater than the critical value for a given degree of freedom and a given level of significance, the test discards $H0$, otherwise it retains it. I used a significance level of $\alpha = 0.05$.

5.3.2 Mann-Whitney U test

Two independent sample sets can be compared using the Mann-Whitney U test, which is also a non-parametric procedure [107]. With this test, I will investigate the homogeneity of the two groups in light of their distribution with respect to the severity of dysphonia.

The Mann-Whitney U-test is an alternative to the independent samples t-test, and it does not assume any specific probability distribution. It is not unusual to use the Mann-Whitney U-test when the assumptions of the independent samples t-test have been violated, in particular the assumption of normality. The null hypothesis (H_0) is that the population distributions are the same in the two independent sample sets. The alternative hypothesis (H_1) is that the distributions are not equal in the two sample sets.

The condition in which one can use this test are the following: the sample must be random, the observations within each sample are independent, as well as the sample sets must be independent from one another.

To perform the test, the following steps must be performed: first, the elements of the two sample sets are merged and ranked in order (based on their values), and to each of their values rank numbers are assigned. The sum of the rankings of the groups (R_1 and R_2) is then determined for each group. If the groups do not have the same number of values, denote the smaller group by N_1 and the larger group by N_2. We calculate the U-statistic for the smaller sample using equation 10.

$$U = N_1 \cdot N_2 + \frac{N_1(N_1+1)}{2} - R_1. \tag{10}$$

The mean ($\bar{x}_B$) and the variance (S_U^2) of the distribution are given by equations 11 and 12:

$$\bar{x}_U = \frac{N_1 \cdot N_2}{2}, \tag{11}$$

$$S_U^2 = \frac{N_1 \cdot N_2(N_1+N_2+1)}{12}. \tag{12}$$

If N is greater than 8, then U has an approximately normal distribution. In that case, the standardized value is given by formula (13.

$$z = \frac{U - \bar{x}_U}{S_U}, \tag{13}$$

where S_U is the standard deviation of the distribution.

Based on the z value, we can decide on the H_0 (H_0: $R_1 = R_2$) hypothesis. If, at the 0.05 level the relation -1.96 $\leqslant z \leqslant$ 1.96 is true, then hypothesis H_0 is accepted, otherwise it is rejected. Rejection of H_0 represents a significant difference in the distribution of samples (H_1: $R_1 \neq R_2$).

5.3.3 Pearson correlation

In statistics, by correlation we typically mean the linear relationship between two probability variables [108]. If two probability variables are independent, they are also uncorrelated, but the opposite is generally not true, although in the case of normally distributed probability variables, uncorrelation also indicates independence. If the correlation coefficient is 0 (H0), then the two variables are uncorrelated, otherwise there is a linear relationship (H1). Acceptance or rejection of H0 is determined by the correlation coefficient as follows:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}, \quad df = n-2, \tag{14}$$

where *df* is the degree of freedom, n is the number of sample pairs of the two probability variables, and r is the correlation coefficient calculated by the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{15}$$

where x_i and y_i are the individual sample points indexed with i, $\bar{x}$ and $\bar{y}$ represent the sample means. If t is greater than the critical value for a given degree of freedom and a given level of significance, the test discards H_0, otherwise it retains it.

To assign a meaning to the strength of the correlation, I used the guide that Evans suggests to interpret the absolute value of r [109]:

- 0.00-0.19 “very weak”;
- 0.20-0.39 “weak”;
- 0.40-0.59 “moderate”;
- 0.60-0.79 “strong”;
- 0.80-1.0 “very strong”;

To inspect the relationship of the acoustic features with the severity of dysphonia the Pearson product-moment correlation coefficient was calculated, as defined in equation 15.

To determine whether the correlation between the variables is significant, one must compare the p-value to a significance level. During correlation analysis $\alpha = 0.01$ level was used. This significance level indicates that the risk of concluding an existing correlation, even if it doesn't exists, is 1%. When performing correlation analysis, if the calculated p-value is less than or equal to the significance level, we can conclude that the correlation is different from 0. If the p-value is greater than the significance level, we cannot conclude that the correlation is different from 0.

5.3.4 Reliability Analysis

During my research I examined the consistency of four specialists' RBH ratings with Cronbach's Alpha and the Intra Class Correlation Coefficient (ICC). Both methods are widely used to estimate the reliability of a composite score.

Cronbach's alpha

Cronbach's alpha is the most commonly used metric for measuring and expressing internal consistency. Statistics are mainly used to test the consistency of respondents in questionnaires, i.e. the reliability of their answers [110]. Sample size can influence Cronbach's alpha results scores, that have a low number of items tend to have lower reliability.

To use Cronbach's alpha the data must satisfy the following conditions: unidimensionality, (essential) tau-equivalence and independence between errors. Unidimensionality assumes the questions we are measuring have only one dimension [111]. The tau-equivalence assumption means that the same true score applies for all test items, or equal factor loadings of all items in a factorial model [112].

The relationship between the alpha value and consistency is shown in Table 4.

Table 4: Meaning of Cronbach's alpha values.

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$\alpha > 0.5$	Unacceptable

The alpha parameter is calculated as follows:

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}, \tag{16}$$

where N is equal to the number of cases, $\bar{c}$ is the mean covariance of the case pairs, here, the mean covariance of the four estimates, and $\bar{v}$ is the average variance over the the four estimates.

Intraclass correlation coefficient (ICC)

In statistics, intraclass correlation coefficient is an inferential statistic that can be used as a measure of the reliability of measurements or ratings. [113]. It describes how strongly the given results of each measurement are similar to each other. The assumptions of ICC include normality and homogeneous variance [114].

Its calculation follows the below principle:

$$r = \frac{MSB - MSW}{MSB + (k-1) \cdot MSW}, \tag{17}$$

where MSB (Mean Square Between groups) is the mean square among sample means, while MSW (Mean Square Within groups) is the within group variation divided by its degrees of freedom, and k is the number of cases.

Cicchetti in [115] provides the following interpretation for ICC inter-rater agreement measures (Table 5):

Table 5: Meaning of ICC r values.

ICC	Internal consistency
r <0.4	poor
0.4 <r <0.59	fair
0.6 <r <0.74	good
0.75 <r <1	excellent

5.4 Classification and regression methods

Applying machine learning methods, I performed classification and regression tasks to investigate the accuracy of the automatic detection of dysphonia and the severity of dysphonia using acoustic-phonetic features or phone level posterior probability values derived from speech. During my research several machine learning methods were tested in order to perform these tasks. In this Section I will briefly describe each method.

Binary classifications

The goal of binary classification was to distinguish healthy voices from the ones with dysphonia in the case of adults and children. Furthermore, the automatic separation of functional and organic dysphonia was performed as well.

For classification tasks I used support vector machines (SVMs) and Deep neural networks (DNNs). The advantage of SVM is that it works well on small datasets. SVMs do not only learns the training points, but also achieve high generalization ability. It provides particularly good results for binary classifiers. DNNs on the other hand require "relatively" large datasets to work well.

The best results for distinguishing healthy voices from the ones with dysphonia were obtained using a *Fully-Connected Deep Neural Network*, with 4 hidden layers, each of them

with 25 neurons [116]. ReLU (Rectified Linear Unit) activation function was used on the hidden layers, Softmax on the output layer. The computationally efficient Adam optimizer was used to update the network's weights during training. This method can be used instead of the classical stochastic gradient descent procedure [117]. To avoid overfitting, I used dropout (with value of 0.25). More on it can be found in Section 5.4.2 and 6.2.

For classification with support vector machines, I used C-SVM with linear and radial basis function (rbf) kernel [118] [119]. SVM is a supervised machine learning algorithm which is used mainly for binary classification tasks [118]. It uses the kernel trick to transform data and based on these transformations it finds an optimal boundary between the classes. The classifier was used successfully in our previous work achieving high accuracy separating the healthy and pathological voices [69]. Since DNNs require a lot of data, the classification of functional and organic dysphonia and the classification of healthy and children with dysphonia was performed with SVM.

Unsupervised cluster analysis

The *k-means* is one of the simplest clustering approaches [120]. This method is a fast and simple approach to the problem: it is easy to implement, and it is easy to interpret the clustering results. Cluster analysis is used to classify cases into relative groups called clusters, in this case: individual assessments of severity of dysphonia. In cluster analysis, there is no prior information about the cluster membership for any of the data. If the acoustic feature set and the unsupervised learning method are fixed, it is possible to compare four cluster models, in each case labelled by a specialist's judgement. In order to examine the subjective nature of RBH k-means, cluster analysis was done. More on this method in Section 5.4.1.

Regression analysis

Support vector regression (SVR) with linear and rbf kernel was used in order to automatically determine the severity of dysphonia [121]. By its nature, linear regression only looks at linear relationships between dependent and independent variables; linear regression also assumes that there is a straight-line relationship between the input variables and the target. SVR with rbf kernel has good generalization and strong tolerance to input noise. For the regression task epsilon-SVR was used [121].

Optimization techniques

The hyperparameters of the SVM and SVR methods were selected by *grid search*. In case of grid search, you can specify which values to test for each hyperparameter, and the function will automatically run the training for each given case and specify which combination achieves the best result. The advantage of this method is that it is simple, it tries out each given combination by itself, without having to manually adjust it. On the contrary, it only works with discrete variables, and because it tests every combination, it can be very slow and time consuming to have a slightly larger set of parameters to try.

In case of SVM and SVR with radial basis function hyperparameters (C - cost and γ) were tested with grid search with values of the power of 2 in range -10 to 10.

In case of the DNN, I did not use an automatic tool to determine the best hyperparameters, instead, I tried out several combinations of the number of hidden layers, the number of neurons in the hidden layers and the droupout values, following rules of thumb. In my experiments, I modified the number of hidden layers from 2 to 5. I adjusted the dropout value between 0.2 and 0.5 in increments of 0.05. I tested the number of neurons in the hidden layers between 10 and 50, with an increase of 5 neurons. The configuration that yielded the best results, is the one with 4 hidden layers, containing 25 neurons in each layer and a dropout value of 0.25. Consequently, that is the one referenced whenever I mention the DNN model later in my Thesis.

5.4.1 k-means clustering

In unsupervised learning the data have no target attribute. What we are interested in is discovering the structure underlying the data. We want to explore the data to find some intrinsic structures in them.

Clustering is a technique for finding similarity groups in data, called *clusters*. The goal is to show that there are groups of data sets that are more similar to each other than members of other groups.

Clustering is often synonymous with unsupervised learning. This is because it takes no class values denoting an a priori grouping of the data instances, as opposed to what we see in supervised learning.

A cluster is a collection of data items which are ‘similar’ between them, and ‘dissimilar’ to data items in other clusters. The data within each cluster is similar to each other on

the basis of some dimension and attributes, and at the same time they are different from the elements of the other clusters. The formation of groups, i.e. clusters, can be done on the basis of different distances and / or similarity measures. Another important feature of the process is that the goal is to create groups that do not overlap as much as possible. Note, some methods (fuzzy clusters) allow the creation of overlapping clusters, which may be needed in specific applications.

The fact that two elements belong to the same group does not mean that they are the same in every respect, similarity of certain attributes may be sufficient. The procedures used in practice fall into two broad categories: hierarchical methods and non-hierarchical methods.

K-means is one of the simplest algorithms that uses unsupervised learning method to solve known clustering issues [120]. It assigns each element to the cluster whose center is closest to that element. This method is a fast and simple approach to the problem: it is easy to implement and the clustering results are easy to interpret. The standard algorithm (also referred to as naive k-means) for a given an initial set of k means $m_1^{(1)}, ..., m_k^{(1)}$ alternates between two steps: the assignment step and the update step [122].

In the assignment step we assign each observation x_p to the cluster $S^{(t)}$ with the least squared Euclidean distance:

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\}, \tag{18}$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more clusters.

In the update step the means (centroids) are recalculated for the observations assigned to each cluster with the following formula:

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j. \tag{19}$$

The biggest advantage of the algorithm is its simplicity and speed, which allows it to be used on large data blocks [123]. The disadvantage, however, is that it does not give the same result after different runs because the result of clustering is affected by the initial randomization. It minimizes the variance within clusters, but does not result in the lowest variance overall.

5.4.2 Neural Networks

Deep learning is part of a broader family of machine learning that performs complex calculations based on artificial neural networks with representation learning. Nowadays, it is gaining ground and is getting better results in many applications. In this subsection I cover the basics of Artificial Neural Networks (ANN).

The neuron

The most popular neuron model is the "perceptron". The perceptron is a neuron capable of separating two sets of linearly separable samples after training. In essence, it implements a binary classifier that is made up of a neuron. In this context they are already called elemental neurons. It was named after the neurons in the brain, because their structure and function were modelled by them. Figure 3 shows the structure of the neuron. It assigns a weight to each of the n inputs, then adds up the product of the inputs and their respective weights. It then applies a so-called activation function to the sum that ensures nonlinearity [116]. In addition to the weighted inputs, an extra input, called bias is usually added with a value of 1, which ensures that you may get a sum other than 0 even if each input sample has a value of 0.

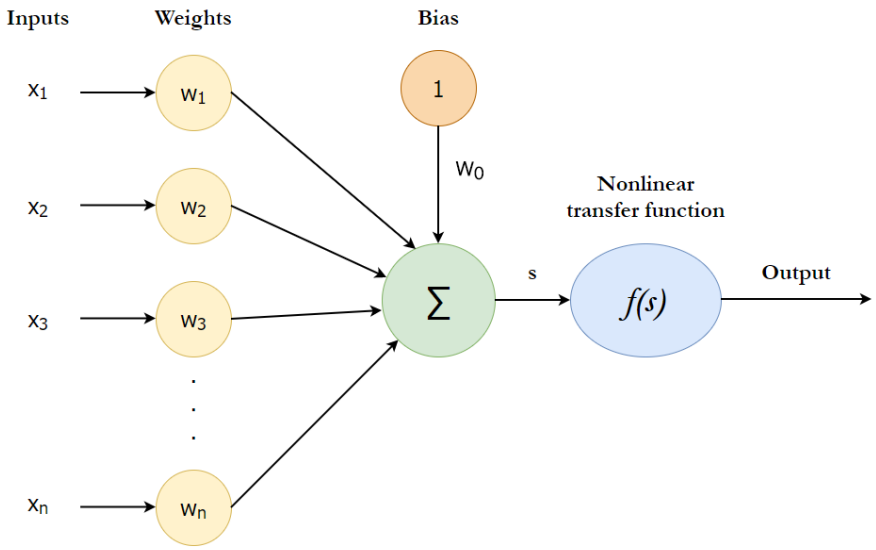

Figure 3: Artificial neuron.

Activation functions

The activation function usually provides nonlinearity for neural networks, which is needed because most of the problems and data in the world are nonlinear. When applied to the sum of the inputs and weights multiplied by elementary neurons, it gives the output. The interpretation range of the output is usually a set of real numbers. The set of values of the output is a finite subset of real numbers, for example, a range of 0 to 1, or -1 to 1.

There are many activation functions, the choice of which depends on the problem to be solved and the available data. The most common ones today include Linear, Sigmoid, Reflected Linear Unit (ReLU), Tangential Hyperbolic (TanH) and Softmax activation functions. The Linear Activation function is the simplest, in fact, it simply returns the value obtained. TanH and Sigmoid are activation functions shown in Figure 4 and Softmax in Figure 5.

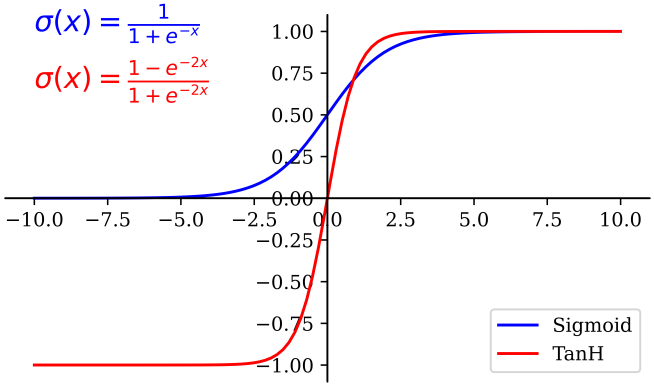

Figure 4: Sigmoid and tangent hyperbolic activation functions.

In the equation in Figure 5, x is the input vector and K is the number of classes. The Softmax activation function is generally used for multi class single label classification, so usually it is used as an output node activation function. Softmax outputs a probability distribution. If we sum the probabilities of all the classes, we get one as a result. The predicted class is the one having the highest probability output.

Artificial Neural Network - ANN

The Artificial Neural Network (ANN) is one of the main tools used in machine learning, often referred to simply as a "neural network". The network itself is not an algorithm, but

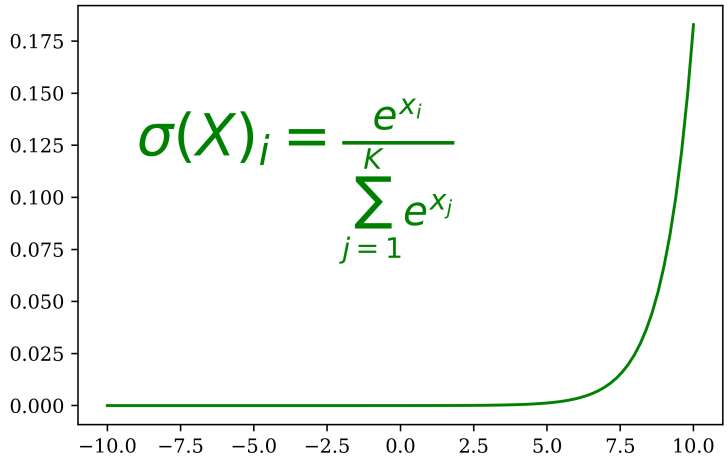

Figure 5: Softmax activation function.

rather a framework for machine learning algorithms [124]. In its name, the "neural" part implies that the system's operation was inspired by the brain and its function, and with the help of it, we are trying to replicate the way in which a human being is learning. Like the brain, ANN is made up of nodes, that is, neurons and the connections between them. The network consists of an input and output layer, and possible hidden layers between the two, in which neurons are grouped. The neurons in each layer transmit the signals from their input to the next layer with the help of their connections. They pass the processed signal until the output layer. The connections are assigned to a weight by which the passing signals can be weakened or amplified. These weights are usually real numbers. As a result, by the time the input data passes through each layer to the output layer, it is transformed.

Based on the connections between individual neurons, we can divide networks into two main groups; feedforward neural networks and recurrent neural networks. If a directed graph representing a network topology contains a loop, it is called recurrent, otherwise it is feedforward.

Deep Neural Network - DNN

Numerous variants of neural networks already exist for various problems with different structures, such as deep neural networks that have many hidden layers, convolutional neural networks that work on the convolution of input data and shared weights, or recurrent neural networks that already have memory and this makes them suitable for time series tasks. A deep neural network is a neural network that has many hidden layers. There is no exact

definition of how many hidden layers are considered deep, but usually two or more hidden layers are already categorized as deep. As an example, in Figure 6 shows a Fully-Connected feedforward deep neural network.

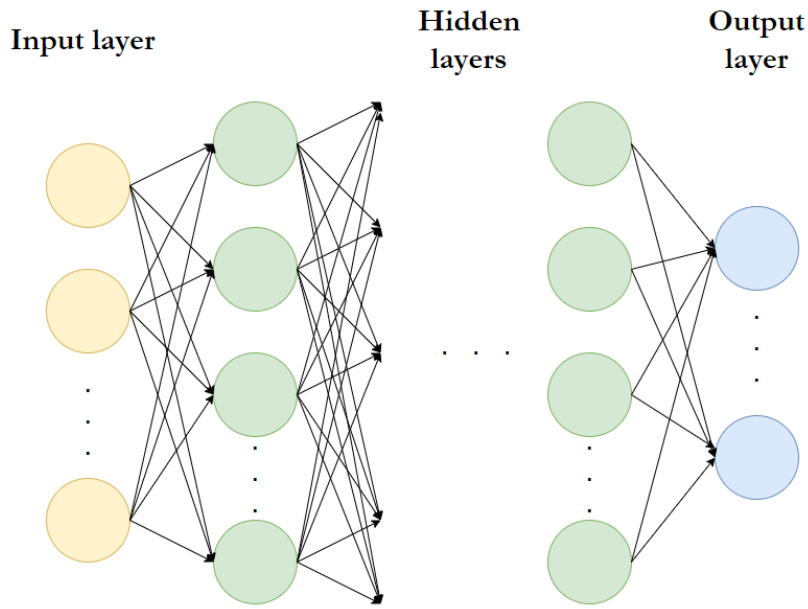

Figure 6: A feedforward deep neural network.

Stochastic Gradient Descent (SGD), cost functions and back propagation

The weight of the connections at first are randomly initialized, but can be trained with Stochastic Gradient Descent (SGD) using backpropagation as a gradient computing technique [125, 126, 127]. The SGD is one of the simplest general optimization algorithms in multivariate functions. The backpropagation algorithm is a fast technique of computing the gradient of the cost function. The training of the neural network with SGD goes as follows.

The training of neural networks consists of two parts: forward-propagation and backpropagation. During the forward pass the data is received by the neurons at the input layers, their outputs are calculated, and then transmitted to the neurons of the next layer through their connections. This continues until the output layer results in a value. The next step is to evaluate the predicted output $\hat{Y}$ against an expected output Y. This evaluation between $\hat{Y}$ and Y happens through a *cost function* (or loss function), which can be a simple MSE (mean squared error) or a more complex function like *cross-entropy*. The lower the loss value, the more accurate is the model, but if the loss is a high value, the training goes in the wrong

direction and the weights need to be optimized. Depending on the nature of the task (e.g., a classification task or a regression task) different cost functions might be worth applying.

The next step is backpropagation, where backpropagation aims to minimize the cost function by updating the network's weights. The level of adjustment is determined by the gradients of the cost function with respect to those parameters. During backpropagation the partial derivative of the error from the cost function (with respect to each parameter) is calculated in a backward pass through the network. The extent to which the algorithm modifies the weights depends on the *learning rate*. The learning rate is used to scale up or down the magnitude of the update. If the learning rate is too low, the search for the minimum loss can be very slow or you may even get stuck in a local minimum. Taking too big steps in the loss space is avoidable as well. If the learning rate is too high, there is a chance that you will not find the minimum. To avoid this, it is a good idea to use optimization algorithms that use an adaptive learning rate.

The gradient descent only works with derivable functions, so it is important to choose appropriate activation functions. In practice, the Tangential Hyperbolic and Sigmoid functions are widespread. The two activation functions have the same shape only their value set is different. While the value set of Tangential Hyperbolic is (-1, 1), the value set of of Sigmoid is (0, 1) (see Figure 4).

Vanishing gradient problem and ReLU

The purpose of the SGD is to determine to what extent the actual value of each weight contributes to the error. As shown in Figure 4, the Sigmoid activation function "squeezes" the values into 0 and 1 and Tangential Hyperbolic activation function between -1 and 1. As a result, when applied to hidden layers, the weights become very low, so that a larger change in the input is hardly perceptible at the output. This is a so-called vanishing gradient problem, which is solved by using the ReLU activation function on hidden layers, or some variation of it, such as Parametric ReLU, as described as shown in Figure 7. The latter is useful because while conventional ReLU resets negative values to 0, parametric ReLU allows values below zero to pass through this range with a parametric multiplier (for example 0.1). All you have to do is enter the parameter, which is 0.1 for the function shown in the figure.

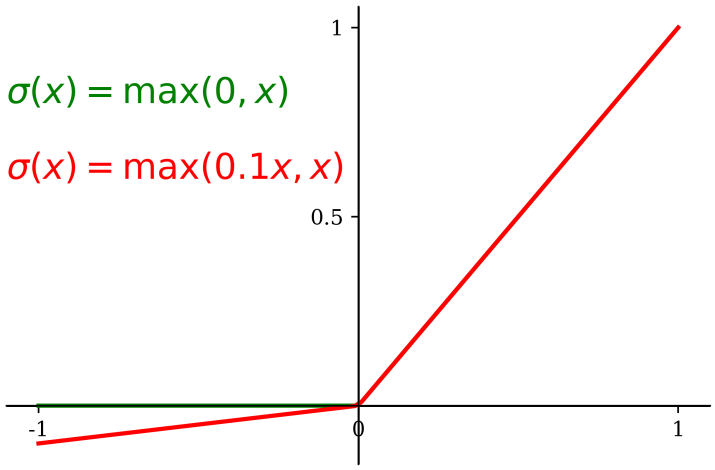

Figure 7: ReLU and Parametric ReLU activation functions.

Regularization techniques

Different regularization techniques can be used to avoid overfitting such as early stopping or L1 and L2 regularization. In my work I used the so-called *dropout* regularization technique [128].

During dropout, we can specify a probability value per neuron so that in some iterations their outputs are not taken into account when calculating the model end result. By doing so, it simulates that multiple models are taught and their results aggregated, thus avoiding overfitting of the training data. It is a generally accepted rule of thumb that the dropout value should be between 20% and 50%.

5.4.3 Support Vector Machines

The purpose of SVM is to create an optimal linear model for classifying samples represented by a feature vector, that is, to determine the hyperplane in the space defined by the feature vector that properly separates the training samples [118]. It is basically suitable for two-class classification, which can be expanded to include multi-class classification. For a two-class classification, the two classes are denoted by -1 and +1.

Linearly separable case

The training dataset samples are said to be linearly separable if you can specify a

hyperplane between the classes defined by the samples that separates the classes without error. However, it is possible that more than one such hyperplane can be specified, in which case the SVM returns a hyperplane with a maximum margin, which maximizes the generalization ability. The hyperplane can be specified by the equation $w \cdot x + b = 0$, where w is a vector perpendicular to the hyperplane, $\frac{b}{||w||}$ is the distance between the hyperplane and the origin.

To determine the hyperplane the SVM selects the x vectors that are closest to the hyperplane from the training vectors, called support vectors. The main task of the SVM is to determine the values of w and b so that the following set of equations is satisfied for the training set while maximizing $\frac{1}{||w||}$:

$$y(x_i \cdot w + b) - 1 \geq 0, \forall i \in I, \tag{20}$$

where y is the representation of the two classes by -1 and + 1, and I is the set of support vector vectors.

Non-linear case

The SVM can also classify non-linearly separable patterns, using the so-called kernel function, that transforms the original problem into a higher dimensional space where it is possible to specify a separating hyperplane. The most common kernel functions are:

$$\begin{aligned}
&Linear: \quad K(\theta(x_i), \theta(x_j)) = x_i^T \cdot x_j, \\
&Polynomial: \quad K(\theta(x_i), \theta(x_j)) = (\gamma \cdot x_i^T \cdot x_j + r)^d, \gamma > 0, \\
&Radial - basisfunction(RBF): \quad K(\theta(x_i), \theta(x_j)) = \exp(-\gamma(||x_i - x_j||)^b), \gamma > 0, \\
&Sigmoid: \quad K(\theta(x_i), \theta(x_j)) = tanh(\gamma \cdot x_i^T \cdot x_j+)
\end{aligned} \tag{21}$$

where d, γ and r are hyperparameters for the given kernel functions.

However, transforming into too high a dimensional space can have a negative effect on the generalization ability of the model, so it is worth introducing a weakening variable ζ_i, which specifies the location of each vector in terms of the decision. $\zeta_i = 0$ indicates that the given vector is well classified and located outside the hyperplane margin. $\zeta_i = 1$ indicates that the given vector is misclassified, and $0 < \zeta_i < 1$ indicates that, although the vector is well classified, it is within the margin of the hyperplane. Based on this, the original equation system is modified as follows:

$$y(\theta(x_i \cdot w + b) - 1 + \zeta_i \geq 0, \forall i \in I, \tag{22}$$

while maximizing $\frac{1}{||w||}$, we can write the optimization problem as follows:

$$min_{w,b,\delta}\{\frac{1}{2}||w||^2 + C\sum_{i=1}^{k}\zeta_i\}, \tag{23}$$

where k is the number of training samples. The constant $C \geq 0$ hyperparameter controls the magnitude of the penalty caused by ζ_i. For $C = 0$, we obtain the linearly separable case.

5.4.4 Support Vector Regression

In SVR, the training samples are not assigned to a class label but to a real number (y). The task of the SVR is to select a function $f(x)$ that approximates the y targets in an optimal way. The optimum mode for SVR means looking for a function $f(x)$ that follows the y values with an error not larger than ε, with no sudden large changes, thus compromising between the accuracy and generalization ability of the function.

The decision function of the SVR can be written in a matter similar to what we saw in Section 5.4.3:

$$f(x) = w \cdot x + b. \tag{24}$$

In order to minimize the change of $f(x)$, the value of w must be minimized. Like SVM, this is equivalent to minimizing $\frac{1}{2} \cdot ||w||$, while satisfying the following system of equations:

$$\begin{cases} y_i - w \cdot x_i - b \leqslant \varepsilon \\ w \cdot x_i + b - y_i \leqslant \varepsilon \end{cases} \quad \forall i \in I, \tag{25}$$

Similarly to SVM, it is possible that the system of equations cannot be solved, i.e. that the function $f(x)$ with an error $\leq \varepsilon$ does not exist in the original space of the training samples. In this case, the method shown in SVM can be applied here too, so errors must be enabled by adding the weakening variables ($\zeta_i \quad and \quad \zeta_i^*$) and transforming the original

problem space using kernel functions. Based on this, the original equation system is modified as follows (while minimizing $\frac{1}{2} \cdot ||w||$):

$$\begin{cases} y_i - w \cdot \theta(x_i) - b \leqslant \varepsilon + \zeta_i \\ w \cdot \theta(x_i) + b - y_i \leqslant \varepsilon + \zeta_i^* \end{cases} \quad \forall i \in I, \tag{26}$$

This means minimizing the following expression while satisfying the former system of equations:

$$min_{w,b,\delta}\{\frac{1}{2}||w||^2 + C\sum_{i=1}^{k}\zeta_i + \zeta_i^*\}, \tag{27}$$

where k is the number of training samples. The constant $C \geq 0$ hyperparameter can be used to the minimize the change of the function.

5.4.5 Model building and testing procedures

Data can have different values by several orders of magnitude. Machine learning algorithms perform better when numerical input variables are scaled to a standard range, so we usually transform the available data, usually between -1 and 1 or 0 and 1, depending on the problem to solve. There are several ways to do this, depending on the nature of the problem, we are working with. We must pay attention to outliers because they can greatly distort the data set during transformation [129].

One of the most commonly used methods is *standardization*, which converts data to have an expected value of 0 and variance of 1. This can be achieved by determining the distribution mean and standard deviation for each feature. In the next step we subtract the mean from each feature, then we divide the values by its standard deviation. The way to perform the standardization (or Z-score normalization) is shown in equation 28.

$$x' = \frac{x - \bar{x}}{\sigma}. \tag{28}$$

In the equation x is the original feature vector, $\bar{x}$ is the mean of that feature vector, and σ is its standard deviation.

Another method is called *min-max scaling*. In this case, the data is scaled to a fixed range, typically between -1 and 1 or 0 and 1. The scaling can be carried out according to

the equations 29 and 30, where x_{min} represents the minimal value of the data and x_{max} represents the maximum value.

$$x_{scaled[0,1]} = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{29}$$

$$x_{scaled[-1,1]} = 2\frac{x - x_{min}}{x_{max} - x_{min}} - 1. \tag{30}$$

During training the values of the acoustic-phonetic features included in the feature vector were normalized between -1 and +1 in the case of SVM and 0 and 1 in case of DNN. In case of regression the target variable (the overall hoarseness H from the RBH scale) values were normalized as well.

To estimate and compare the performance of the machine learning algorithms *Leave-one-out cross validation (LOOCV)* was used, where the result is a large number of performance measures that can be summarized in an effort to give a more reasonable estimate of the accuracy of your model on unseen data. A downside of this approach is that it can be a computationally more expensive than a k-fold cross validation approach.

During LOOCV, it is not necessary to separate the dataset into training and testing sets. A test consists of as many iterations as the number of samples in the dataset. At each iteration, one sample is selected for testing and all others for training. We run the training / testing procedure and select another sample to test and use the rest to train, accordingly. This way, the evaluation process takes much longer, but we can use all the samples in the database for training and testing. The use of this method is necessary due to the nature of my research, that is, it is very difficult to access medical data, therefore a small sample size was available.

Note, some experts suggest the use of *nested cross-validation* (also called double cross-validation) to avoid the optimistic bias resulting from using the same cross-validation to set the values of the hyperparameters of the classifier [130, 131]. In this case two k-fold cross validation loops are performed: the inner loop is used for the model selection, and the outer loop estimates the generalization accuracy. Others suggest that nested cross-validation could just be a waste of computational time in cases where the classifier algorithms have only a limited number of hyperparameters that must be tuned [132]. Furthermore, the use of nested cross-validation is suggested when the dataset is small, a few thousand observations or less. On

an even smaller dataset - a few hundred observations or less - the two k-fold cross-validation loops reduce the number of training samples which leads to loss of information.

5.4.6 Feature selection

In order to reduce the dimensionality of the input vector the *Forward Feature Selection (FFS)* algorithm was used [133]. Forward feature selection is an iterative algorithm, choosing the best feature that improves the performance regarding to a cost or objective function. In each step features are added to a set of already selected features. Here, the features were selected using maximum accuracy as an objective function.

This process is structured as follows:

- The procedure starts from an empty feature vector.
- In each round, it adds an unused feature to the feature vector.
- For each added feature, the performance is estimated using, e.g., a cross-validation.
- Only the attribute giving the highest increase of performance (in my case accuracy) is added to the feature vector.
- Then a new round starts with the modified selection.
- Finally, the procedure returns the vector with the highest accuracy.

A stopping behaviour parameter can be set that specifies when the process should be stopped. In my case the iteration runs as long as there is any increase in performance for 3 speculative rounds. This parameter defines how many rounds will be performed in a row, after the first time the stopping criterion is fulfilled. There are two cases: if the performance increases or if it doesn't. If the performance increases again during the speculative rounds, the selection will be continued. If the performance does not increase, all additionally selected features will be deleted, as if there were no speculative rounds at all. This technique helps avoiding the problem of getting stuck in a local optima.

5.4.7 Evaluation methods

In order to describe the performance of a classification or cluster model I present the *confusion matrices*. From the confusion matrices I calculate the metrics *accuracy*, *recall* and *precision*.

Table 6: 2x2 Confusion matrix.

	Actual: Yes (Healthy)	**Actual: No (Dysphonia)**
Predicted: Yes	TP	FP
Predicted: No	FN	TN

An example of a 2x2 confusion matrix is shown in Table 6. The confusion matrix is a table with four different combinations of predicted and actual values: TP, FP, FN and TN.

The *true positive (TP)* is the number of cases when we predicted positive and it is true, in my case the model predicted that a patient is healthy and the patient is in fact healthy. The interpretation of a *true negative (TN)* is when we predicted negative and it is true, in my case the model predicted that the patient is not healthy (predicted dysphonia) and the patient is in fact not healthy (has dysphonia). The model successfully predicted the non-existence of the condition. A *false positive (FP)* (also called as a *Type I error*) happens when we predicted positive and it is false, so the model predicted that the patient is healthy but in reality has dysphonia. In my work it is very important to minimize this number when I attempt binary classification between healthy and dysphonic voices. Finally, *false negative (FN)* (also called as a *Type II error*) happens when we predicted negative and it is false, so the model predicted that the patient has dysphonia but in fact the patient is healthy. From these values, we calculate the evaluation metrics as follows.

Accuracy

Accuracy is a measure of how likely the model is to find the correct answer, in other words, how often is the classifier correct. Calculation method:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{31}$$

Recall

The recall value gives the value of all the positive classes, how much we predicted correctly. It is calculated with the equation 32.

$$recall = \frac{TP}{TP + FN} \tag{32}$$

Precision

The precision answers the question if the model predicts yes, how many times it is correct. Equation 33 gives us the calculation method.

$$precision = \frac{TP}{TP + FP} \tag{33}$$

In the medical world, it is common to use terms such as *specificity* and *sensitivity* to evaluate medical tests. The concepts are very similar, but it can be misleading. To avoid misunderstandings, I give the definitions of the two terms.

Specificity

Specificity is the true negative rate, or the proportion of true negatives to everything that should have been predicted as negative.

$$specificity = \frac{TN}{TN + FP} \tag{34}$$

Sensitivity

Sensitivity measures the proportion of positives that are predicted correctly. It is calculated with the equation 35.

$$sensitivity = \frac{TP}{TP + FN} \tag{35}$$

We can see that recall and sensitivity are the same, but precision and specificity are different.

To describe the accuracy of the regression tasks, two descriptive features are given. The performance of the regression methods is evaluated by the *root mean square error (RMSE)* value, the linear relationship between the target and the predicted H scores is described by *Pearson correlation*.

Root mean square error

The RMSE is a widely used measure to describe a regression model performance, by measuring difference between predicted values and the actual values. It is calculated by the following equation.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}(\hat{y}_t - y_t)^2}{N}}. \tag{36}$$

In the equation $\hat{y}_t$ are the predicted values and y_t are the actual values, while N is the sample size.

Pearson correlation

Pearson correlation coefficient is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. It is used to describe the effect of change in one variable when the other variable changes. The value of the coefficient is between +1 and -1, where value of +1 is interpreted as perfectly positive correlation, 0 is no linear correlation, and -1 is perfectly negative correlation.

The calculation of the Pearson correlation is shown in equation 15.

ASR phone error rate

The ASR phone classifier DNN soft-max layer generates a 10 ms frame-level phone posterior distribution (as mentioned in Section 5.2.2). We can get the phone with the highest probability from the each 10 ms frame, thus obtaining a phone label for each 10 ms. The obtained phone label can be compared with a reference (given by the automatic phone segmentator's forced alignment and the manual corrections) frame by frame. We have two scenarios: we get a matched frame if the obtained phone is the same as the reference. If the obtained phone is not equal to the reference we count it as a mismatch and the phone error rate (PER) is computed as follows:

$$PER = \frac{\text{Total number of phones mismatched}}{\text{Total number of phones in the reference sequence}} \tag{37}$$

5.4.8 Softwares used

Voice recordings were made using Cool Edit Pro. I needed to mark speech sounds and their boundaries, for example where the vowels are positioned in time in a voice recording. For this task the BME-TMIT-LSA Language-independent Forced Speech Segmenter was used [134]. The manual correction of the segmentation was made in Praat [135]. Acoustic features were also calculated with the help of Praat.

For SVM SVR classifications and regression tasks LibSVM was used [136]. The DNN used for classification was implemented in an open-source neural-network library written in Python [137] called Keras [138].

To carry out statistical tests IBM SPSS Version 22.0. was used [139].

Data preparation, classifications and regression tasks was done in RapidMiner [140] and with Python [137].

Figures and plots were also created with Python. The examples of the EMD decomposition were made with the package PyEMD [141].

6 Results

6.1 The examination of the automatic assessment of the severity of dysphonia

6.1.1 Phonetic-class based correlation analysis for the severity of dysphonia

In the diagnosis and management of dysphonic speech, a voice clinician typically assesses the quality of a patient's voice personally. The assessment is subjective by nature. The target severity of a voice is usually defined as one clinician's assessment or as the median or average severity rating determined by a group of experienced raters assessing the voice [73, 142]. If multiple raters are recruited for the subjective assessment of severity of dysphonia, the assessment is done by listening to the previously recorded voice samples. The assessment can vary among raters; thus, analysis of rating consistency is advisable. In the work of Law and his colleagues [143], it was found that higher intra-rater reliability was achieved with continuous speech than with sustained vowel samples. In most voice clinics, acoustic measures are derived from sustained vowel samples; however, continuous speech has several advantages over analysis of sustained vowels. It contains a variation of fundamental frequency, pauses and phonation onsets, and there is the opportunity to examine different combinations of speech sounds.

An important task is to identify relevant acoustic features to predict the severity of the dysphonic voices automatically. The following theses address this issue.

Using a small speech database, it is very important to optimize the speech features as much as possible, rather than to use a lot of acoustic features, with the risk of bringing unwanted noise into the system. My hypotheses are the followings: speech defect severity determined by a clinician (RBH) is correlated (coincides) with the distortion degree of the characteristic acoustic features.

In my first thesis I performed correlation analysis between acoustic features (presented in subsection 5.2.1) and the severity of hoarseness given by a specialist. The specialist treated the patient and determined the diagnosis. The specialist directly listened to and evaluated the quality of the patient's speech during the consultations. The Pearson correlation coefficient was calculated in every case where correlation was significant at the 0.01 level (2-tailed) between the acoustic feature and the subjective rating.

The analysis was carried out on a subset of the database presented in subsection 5.1.1.

The distribution of the voice recordings by H used in this experiment is shown in Table 7. Note that the recordings with the value H equal to 0 are all recordings from healthy patients. Thus, a total of 136 records from healthy people and 206 records from patients suffering from dysphonia were used.

Table 7: Distribution of healthy and dysphonic speakers in the database, depending on the value of H.

Count	**Value of H**				**Total**
	0	1	2	3	
Male	67	34	20	32	153
Female	69	72	27	21	189
Total	136	106	47	53	342

The results presented on Figure 8 indicate that features such as jitter(ddp), shimmer(dda), Harmonics-to-Noise Ratio (HNR) and "mfcc01" correlate with the severity of dysphonia. In the figures, the absolute values of the Pearson correlations are shown where correlation is significant at the 0.01 level (2-tailed). The greater the absolute value of the correlation coefficient, the stronger the relationship between the acoustic features and the severity of hoarseness. According to Evans suggestions, the correlation of jitter.std.[E] and mfcc01.std.[E] with the severity of hoarseness is "very weak", the correlation of shimmer.std.[E] is "weak", the correlation of jitter.mean.[E], shimmer.mean.[E], HNR.mean.[E] and HNR.std.[E] is "moderate", while the correlation of mfcc01.mean.[E] with the severity of hoarseness can be considered "strong".

When SPI was measured on phonetic classes, the Pearson correlation coefficients ranged from 0.11 and 0.23, indicates "very weak" and "weak", but significant correlation. The results are shown in Figure 9.

EMD-based IMF entropy frequency band ratios correlate with the severity of dysphonia, as Figure 10 suggests. IMF.std.[E], IMF.std.[Nasal], IMF.std.[VoicedSpirants], IMF.std.[VoicedPlosives] show "weak" correlation, while IMF.mean.[E], IMF.mean.[Nasal], IMF.mean.[HighVowels], IMF.std.[HighVowels], IMF.mean.[LowVowels], IMF.std.[LowVowels], IMF.mean.[VoicedSpirants] and IMF.mean.[VoicedPlosives] show "moderate" correlation with the severity of dysphonia.

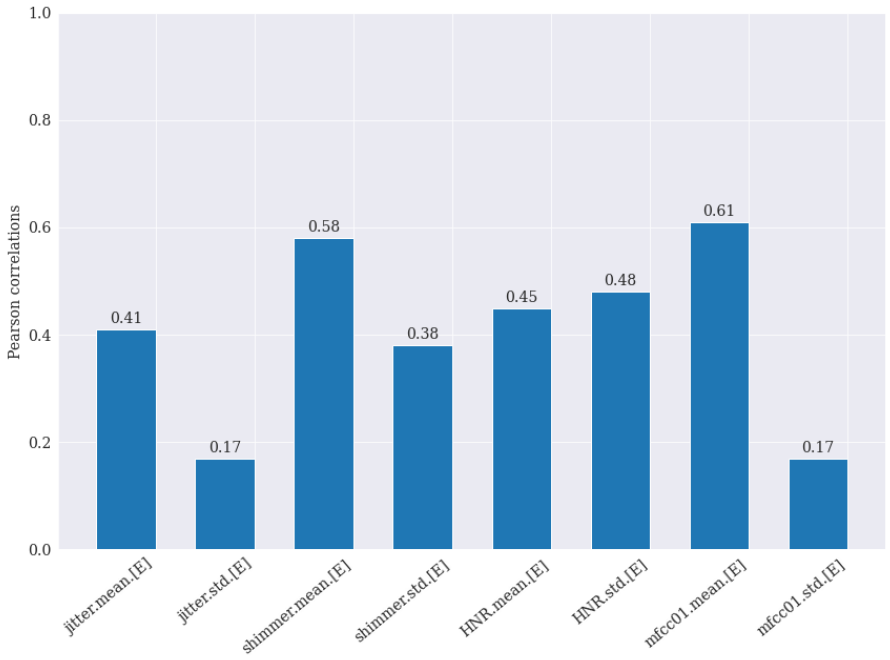

Figure 8: Pearson correlation with commonly used acoustic features.

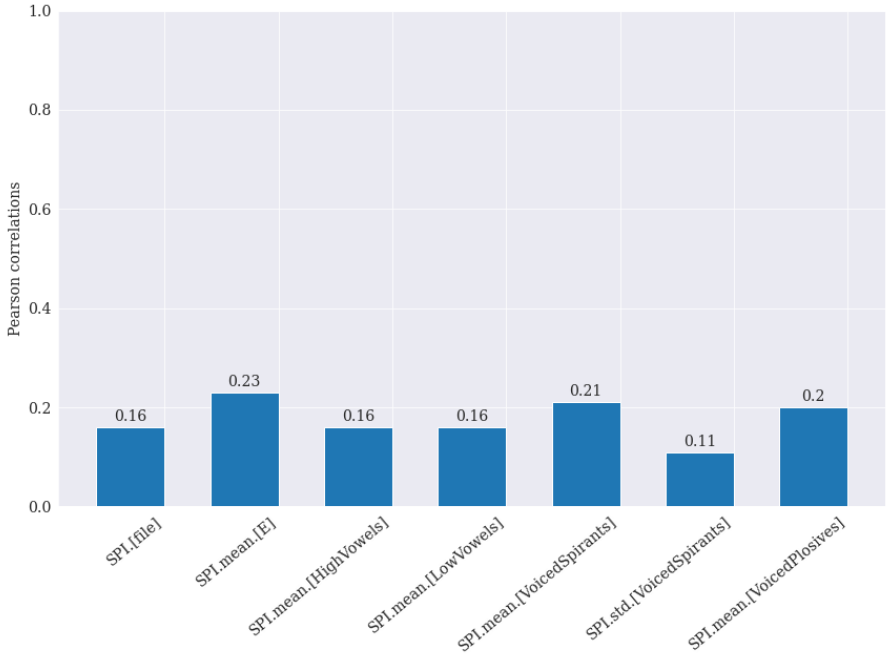

Figure 9: Pearson correlation with SPI measured on phonetic classes.

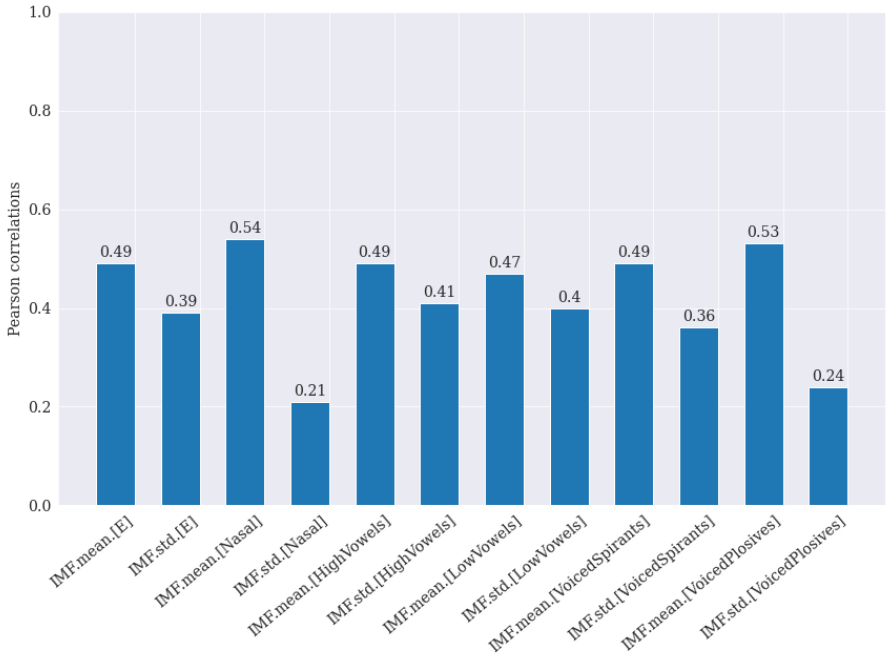

Figure 10: Pearson correlation with IMF entropy-based frequency band ratios measured on phonetic classes.

Thesis I. A. [C4] *I showed that jitter(ddp), shimmer(dda), Harmonics-to-Noise Ratio (HNR), mfcc01, Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based IMF entropy frequency band ratios measured at specific phones show significant correlation at the 0.01 level with the severity of dysphonia when measured on the Hungarian Dysphonic and Healthy Adult Speech Database.*

6.1.2 Unsupervised and supervised learning methods for the modelling of the four grade assessments of the specialists

It is also an important question whether the acoustic features selected by the correlation analysis are suitable for modelling the four grade assessments of the specialists (RBH subjective scale). In this investigation two datasets were used, the Initial Dysphonic and Healthy Database and the Selected Dysphonic and Healthy Database. An unsupervised learning method, the k-means algorithm was used on the Selected Dysphonic and Healthy Database. K-means clustering is a type of unsupervised learning where we have unlabelled data. The goal of the algorithm is to find groups in the data called clusters, with the number of groups represented by the variable k. More details on the k-means method can be

read in Subsection 5.4.1. Before I performed the unsupervised learning method a two-class classification was performed to find out whether the chosen acoustic features are rich enough in information to differentiate between healthy and dysphonic voices, even after reducing the dimensionality of the input vector.

The Initial Dysphonic and Healthy Database contains a total of 263 speech recordings, 127 recordings from healthy subjects (62 male and 65 female) and 136 recordings from patients suffering from functional or organic dysphonia (66 male and 70 female), thus each recording is from a separate subject. The specialist who treated the patient determined the diagnosis. The specialist evaluated the quality of the patient's speech during consultation time. This database was used for the two-class classification experiment.

Four specialists were asked to evaluate the voice recordings of the Selected Dysphonic and Healthy Database with respect to the severity of the dysphonia. The Selected Dysphonic and Healthy Database contains a total of 148 recordings, and it was used for the unsupervised cluster and regression analysis. One of the four specialists set up the diagnosis and evaluated the quality of the patient's speech during the consultations; the other three specialists did not know the patient and only listened to the previously recorded voice files and determined the severity of dysphonia. Every rater is experienced in working with patients with voice disorders and dysphonia.

A two-class classification was performed on the Initial Dysphonic and Healthy Database using leave-one-out cross validation, with SVM classifier. Classification experiments were made using several combinations. Liner and rbf kernels were also tried out. The default value of C of support vector machine is 1, while γ is 1/number of features. In order to choose the optimal hyperparameters for the SVM classifier grid search was used. Leave-one-out cross validation was used in all cases.

The highest accuracy of 89% was reached by using an rbf kernel. The FFS feature selection algorithm reduced the input dimensionality to 18 acoustic features. The acoustic features selected by the FFS algorithm are the following: jitter.mean.[E], shimmer.mean.[E], HNR.mean.[E], mfcc01.mean.[E], jitter.std.[E], shimmer.std.[E], HNR.std.[E], mfcc01.std.[E] ,SPI.std.[E], SPI.mean.[Nasal], SPI.std.[Nasal], SPI.std.[LowVowels], SPI.mean.[VoicedSpirants], SPI.std.[VoicedSpirants], IMF.std.[E], IMF.mean.[Nasal], IMF.mean.[VoicedPlosives], IMF.std.[VoicedPlosives].

It is an interesting question whether the chosen acoustic features can model the individual assessments. Cluster analysis tries to identify structures, homogeneous groups of cases not

previously known within the data. In my case, the hidden structure is the "true label" of severity score for each recording given by an ideal examiner. The cluster analyses is used to mimic this ideal examiner in order to get the true label for each recording. Of course, we rely more on a specialist's rating than on a clustering process, but if the found clusters are really close to a specialist's rating we can consider calling that assessment as true. Hence, I need to compare four cluster models labelled by a specialist's judgement. If the acoustic feature set and the unsupervised learning method are fixed, it is possible to compare four cluster models, for each case, labelled by a specialist's judgement. To examine the subjective nature of RBH, k-means cluster analysis was performed.

k-means clustering has the objective of putting the observations into k clusters, where k is the number of clusters determined by the user as an input. I set the number of clusters to four. The cluster analysis classified the observations into clusters A, B, C, and D.

The clusters where assigned to the severity value by the minimum mean of absolute errors (minimum of MAE): A to H = 0, B to H = 1, C to H = 2, and D to H = 3. MAE was calculated in the following way: initially, the clusters where assigned to the severity values trying all possible combinations. A specialist's judgement represents the true severity value, while the cluster assigned severity is the given severity. The given severity values were subtracted from the true values for each recording, then I took their absolute value. After that I calculated the means in each combination. This is summarized in equation 38.

$$MAE = Mean(|\text{True severity - Given severity}|). \tag{38}$$

The final assignment is where the MAE is minimal.

The confusion matrices for each specialist are shown separately in Table 8, 9, 10 and 11. The accuracies for the decision in case of each specialist in order is: 49%, 44%, 45%, 47%, the mean accuracy is 46.25% with 2.22% standard deviation. In the case of a balanced distribution of 4 classes, the baseline classification would be 25%. From this experiment I can conclude that the acoustic feature set is suitable for modelling the individual assessments of dysphonia severity.

Table 8: Confusion matrix based on the assessment of Specialist 1.

		Specialist 1 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**12**	1	2	1	75%
	1	13	**33**	5	3	61%
	2	9	25	**10**	5	20%
	3	2	4	6	**17**	59%
Class recall		33%	52%	43%	65%	

Table 9: Confusion matrix based on the assessment of Specialist 2.

		Specialist 2 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**11**	3	2	0	69%
	1	5	**26**	23	0	48%
	2	6	24	**16**	3	33%
	3	0	2	15	**12**	41%
Class recall		50%	47%	29%	80%	

Table 10: Confusion matrix based on the assessment of Specialist 3.

		Specialist 3 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**11**	2	2	1	69%
	1	2	**20**	25	7	37%
	2	3	15	**16**	15	33%
	3	0	1	9	**19**	66%
Class recall		69%	53%	31%	45%	

Table 11: Confusion matrix based on the assessment of Specialist 4.

		Specialist 4 (True Label of H)				
		0	1	2	3	Class precision
Predicted label	0	**12**	2	2	0	75%
	1	7	**24**	18	5	44%
	2	6	18	**17**	8	35%
	3	0	6	6	**17**	59%
Class recall		48%	48%	40%	57%	

I calculated the Pearson correlation between the cluster defined severity scores and the individual specialists' ratings and I also evaluated this using the mean RBH perceptual evaluation of the four specialists. Values are shown in Table 12. All correlations show "moderate" relations. The mean correlation is 0.52 with 0.01 standard deviation. The highest value was measured between the cluster defined severity scores and the mean of the ratings, giving a value of 0.59. Since the found clusters correlate the best with the mean of the four specialists, this is the true label I use for the regression analyses.

Table 12: Pearson correlation between the cluster defined severity scores and the specialists' ratings.

	Specialist 1	Specialist 2	Specialist 3	Specialist 4	The mean of the ratings
Pearson correlation	0.51	0.54	0.53	0.51	0.59

Thesis I. B. [C4, J4] *I showed that when clustering the data, with the selected acoustic features using k-means clustering, the found clusters correlate well with the severity of dysphonia. A 0.59 Pearson correlation was achieved between the cluster defined values and the mean of the four specialists' ratings.*

6.1.3 The automatic assessment of the severity of dysphonia with regression analysis

This analysis was performed on the Selected Dysphonic and Healthy Database. As a result of subsection 6.1.2 the mean RBH perceptual evaluation of specialists was used as the target for my regression models.

It is important to analyse whether the rater reliability of the 4 experts is consistent enough. The value of the internal consistency of the specialists gives us an idea of what is the maximum correlation value we can expect from our regression model. We do not expect the regression model to achieve better results than a well trained specialist. For measuring internal consistency ("reliability") of the raters' evaluations Cronbach's Alpha and the Intra Class Correlation Coefficient (ICC) methods were used. Despite the interesting differences among the decision of the specialists, a high degree of reliability (Cronbach's Alpha = 0.89, ICC = 0.89) was measured between their severity judgements when measuring internal consistency.

Regression has a significant advantage compared to cluster analysis, since it's prediction is not an ordinal, but a continuous variable. This property can significantly improve the quality of the model. Due to the small sample size, leave-one-out cross validation was used. The performance of the regression methods is evaluated by the RMSE value, the linear relationship between the target and the predicted H scores is described by Pearson correlation. To find the optimal hyperparameters grid search was used.

In this analysis, support vector regression with linear and radial basis function kernel were used. To reach the best performance the 18-feature set (described in 6.1.2) and the result of the FFS algorithm was used, for SVR with linear and rbf kernel separately. As previously mentioned the mean of the four specialists' ratings was used as target. Table 13 summarizes the results.

Table 13: Regression analysis results – the mean of the four specialist's ratings as target.

Acoustic feature set	**Type of regression**	**Correlation**	**RMSE of H**	**hyperparameters**
18 feature set	linear kernel	0.83	0.50	C = 1
Result of FFS, 8 feature set	linear kernel	**0.85**	**0.46**	C = 1
18 feature set	rbf kernel	0.81	0.51	C = 2, $\gamma = 0.125$
Result of FFS, 14 feature set	rbf kernel	**0.85**	**0.45**	C = 4, $\gamma = 0.25$

The FFS algorithm reduced the original 49-dimension input to only eight features using linear kernel. The following features were selected mfcc01.mean.[E], shimmer.mean.[E], SPI.std.[LowVowels], HNR.std.[E], SPI.mean.[HighVowels], IMF.mean.[Nasal], SPI.std.[VoicedPlosives], IMF.std.[LowVowels]. This configuration gave the highest 0.85 correlation.

When rbf kernel was used, the FSS algorithm selected 14 features, these were the following: shimmer.mean.[E], HNR.mean.[E], mfcc01.mean.[E], HNR.std.[E], SPI.mean.[E], SPI.std.[E], SPI.std.[Nasal], SPI.mean.[HighVowels], SPI.mean.[LowVowels], SPI.std.[LowVowels], SPI.mean.[VoicedPlosives], IMF.mean.[Nasal], IMF.mean.[VoicedPlosives], IMF.std.[VoicedPlosives]. The lowest RMSE value of 0.45 was obtained here. Furthermore, the FFS models gave only slightly better results than the models with the 18-feature set.

This illustrates the capacity of the proposed approach in predicting the severity of dysphonia regardless of the speaker's pathology or severity degree. Since the ICC value

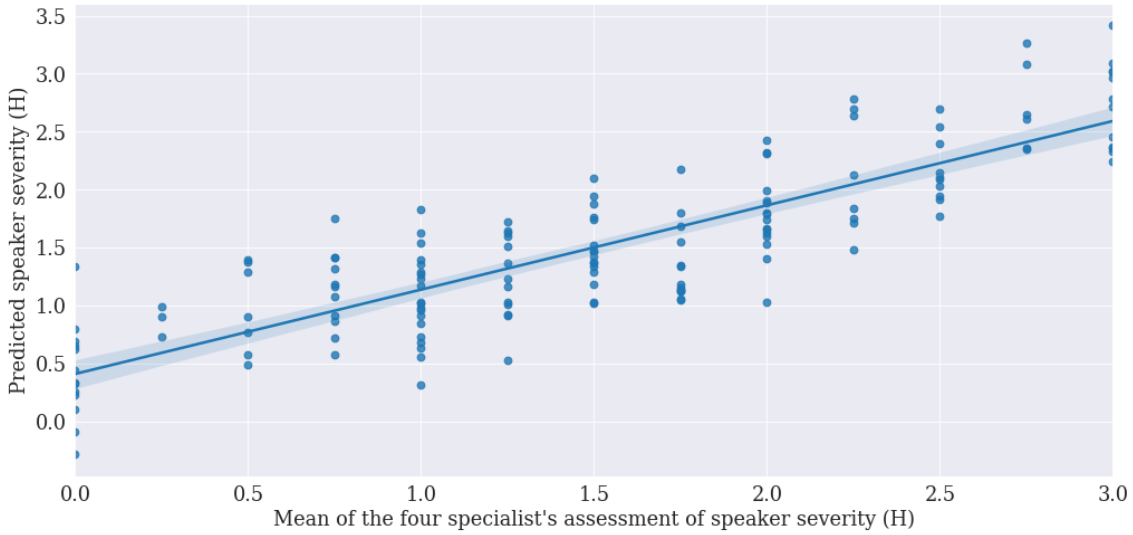

Figure 11: Automatically predicted dysphonia severity degree according to perceptual assessment of H, using SVR with linear kernel regression with 8 parameters.

between the 4 specialists resulted in 0.89 it can be considered as a theoretical goal we want to achieve. In light of this, the correlation value of 0.85 obtained by the regression model is considered almost perfect.

Figure 11 depicts the automatically predicted severity of the dysphonia compared to the reference perceptual assessment of speaker severity. The figure shows the SVR linear kernel regression model created by the result of the FFS algorithm. The figure illustrates once again the capacity of the proposed approach in predicting the severity of dysphonia regardless of the speaker's pathology or severity degree. It can be observed that the model gives good prediction of severity of H = 1.

Figure 12 shows the distribution of the predicted H values from the SVR regression model with linear kernel and the mean of the four specialist's ratings. The distribution of the predicted values of the model is very similar to the H values' mean given by the specialists, resulting in the same means (1.5 and 1.5), but the predicted values have a lower standard deviation (0.86 and 0.73).

Thesis I. C. [C4, J4] *I showed that an automatic estimation of the severity of dysphonia is possible using only eight acoustic features as input vector with a SVR with linear kernel reaching 0.85 Pearson correlation and 0.46 RMSE on the Selected Dysphonic and Healthy Database.*

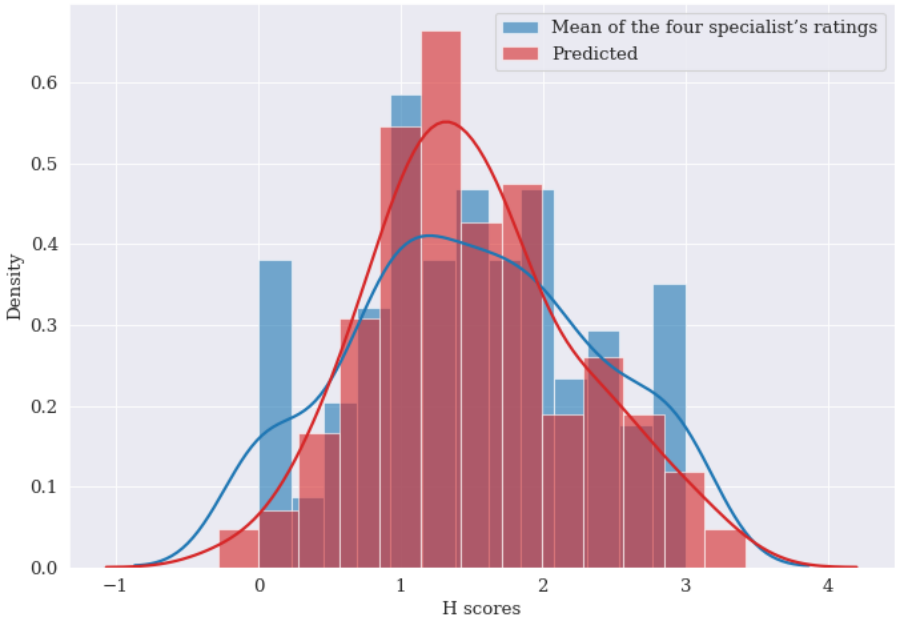

Figure 12: Histogram of the mean of the four specialist's ratings and the predicted H scores.

6.2 The automatic classification of dysphonic and healthy speech

6.2.1 The comparison of SVM and DNN classifiers using acoustic features as an input vector

In order to do a binary classification of dysphonic and healthy speech, researchers generally use a wide variety of acoustic features, derived from speech and used as input vectors with machine learning algorithms [144, 145].

For classification tasks, a common machine learning algorithm is based on SVMs [69, 58], as they are good at dealing with small data samples, but Deep Learning technics are also exploited [146, 147, 148, 149, 150, 151]. Deep neural networks (DNNs) are used on a variety of tasks, usually on big datasets.

In this experiment I used two classification approaches. The first classifier used was SVM, the second classifier was a Fully-Connected Deep Neural Network as described in Section 5.4. FFS algorithm was used in order to reduce dimensionality of the input vector in the case when SVM was used as a classifier. More on the FFS algorithm in Section 5.4.6. In order to choose the optimal hyperparameters for the SVM classifier grid search was used.

The database used in this experiment is the same database described in Section 5.1.1 and in Table 2. The database contains a total of 450 recordings, 257 from patients with dysphonia (156 females and 101 males) and 193 people with a healthy voice (108 females

and 85 males).

I created the input vector from acoustic features described in Section 5.2.1, thus 49 acoustic features were used as input vector.

The results of the binary classification with LOOCV between healthy and dysphonic voices using acoustic features as input vector are shown in Table 14.

Table 14: Two-class classification results between HC and Dys in case of leave-one-out cross validation.

Input vector	**FFS**	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
Acoustic features	Yes	11	SVM	linear kernel	C = 1	85%
Acoustic features	Yes	9	SVM	rbf kernel	C = 256; $\gamma = 0.0625$	85%
Acoustic features	No	49	SVM	linear kernel	C = 4	83%
Acoustic features	No	49	SVM	rbf kernel	C = 1024; $\gamma = 0.00098$	84%
Acoustic features	No	49	DNN	dropout	0.25	**88%**

The first column of the table shows the type of the input vector, the next whether FFS was performed on the input vector or not, then the classifier and the configuration used, followed by the accuracy. In case of SVM, grid search was used in every case. In case of DNN dropout value of 0.25 was used.

Using DNN as a classifier yields higher accuracy than the SVM approach with 3.53% relative accuracy increase, resulting in the highest accuracy of 88%. Also, there is no considerable difference in accuracy between linear and rbf kernel in case of SVM (85% and 85%).

The confusion matrix using FFS and SVM with linear kernel is shown in Table 15, while the confusion matrix of the DNN scenario in Table 16. When using SVM the class precision of the HC class is 84.83%, while the precision of the Dys class is 84.56%. The recall for class HC is 78.24% and 89.49% for class Dys.

The confusion matrix of the 88% accuracy setting is shown in Table 16. As the table suggests, the class precision of the HC class is 79.75%, while the precision of the Dys class is 98.12%. This means that the number of cases where the two classes were predicted correctly

Table 15: Confusion matrix using FFS and SVM with linear kernel.

	true HC	**true Dys**	class precision
pred. HC	151	27	84.83%
pred. Dys	42	230	84.56%
class recall	78.24%	89.49%	

Table 16: Confusion matrix using a Fully-Connected Deep Neural Network.

	true HC	**true Dys**	class precision
pred. HC	189	48	79.75%
pred. Dys	4	209	98.12%
class recall	97.93%	81.32%	

is not balanced. The recall for class HC is 97.93% and 81.32% for class Dys.

Accuracy is not the only absolute measure by which we characterize our classifier. It obscures a lot of important information, so it should be handled with care. We like confusion matrices if they are symmetric, if the mismatch weights of the classes are even. In medical applications, in general, the confusion matrix is not a symmetric matrix. Classifying a sick person as healthy is a more serious mistake than classifying a healthy person as sick. This means that the recall (also known as sensitivity) of class Dys is also a very important aspect of the classifier. The lower the risk that a person with dysphonia is miss-classified as healthy the better. We rather have some healthy people labelled dysphonic over predicting a dysphonic person healthy. In this sense, using FFS and SVM with linear kernel seems to be a better approach since the recall of Dys is 89.49%, while when using DNN the recall of Dys is 81.32%.

If false negatives and false positives have similar costs two decision tables can be constructed by expected number of good predictions and the number of miss-predictions and the achieved number of good predictions and the number of miss-predictions, then chi-square test be performed. The p-value of the test was 0.09, so there was no statistical difference found between the system with 85% accuracy (provided by the SVM with linear and rbf kernel) and the 88% accuracy Fully-Connected Deep Neural Network.

Thesis II. A. [C2, C9] *I showed that the binary classification of dysphonic and healthy voices is possible for Hungarian. When applying a Fully-Connected Deep Neural Network, an accuracy of 88% can be achieved with LOOCV, using acoustic features as input on the Hungarian Dysphonic and Healthy Adult Speech Database.*

6.2.2 Using ASR posterior probability features as input vectors for the DNN classifier

Automatic speech recognition (ASR) is traditionally decomposed into creating an acoustic and a language model with a vocabulary [152]. The acoustic model is most often a hybrid of a Hidden Markov Model (HMM) to facilitate dynamic warping for alignment, and a set of phone or phone alike models (obtained through a decision tree to group acoustically similar entities) responsible for providing similarity measures between the actual frame(s) to be classified and the phone (senone) set. Although this set of models rarely corresponds to pure phone models, for simplicity I will refer to this as a 'phone model', especially as the output of the phone model can be collapsed to phone posteriors. It is obvious that these phone posteriors can be used individually to classify frames, or better, to derive a Goodness of Pronunciation (GOP) score [153] which can be used in speech assessment to automatically evaluate pronunciation.

Using the GOP or pure phone posteriors as additional or standalone features to detect or classify voice disorders – although not particularly dysphonia – has been addressed by many researchers. For example, in [154] and [64] researchers use ASR posteriors to predict the severity of 'general' voice disorders, where the type and characteristics of the disorders are not classified, but their severity is known. The frame level posteriors, produced by a DNN phone model, are a good measure of the acoustic mismatch caused by voice quality change, and thus can be exploited for classification and assessment of voice disorders.

Nevertheless, training a phone model for ASR has a different objective than the recognition of dysphonia requires. An ASR has to tolerate high inter- and intra-speaker variance, and whether samples from dysphonic speech are used or not for its training is not controlled. In brief, a phone model is not trained to discriminate between dysphonia and normal speech. However, it is still used, because of the hypothesis that dysphonic speech is nonstandard and therefore the phone posterior distribution will not be peaky, but rather flat (indicating that the phone classifier is uncertain about its decision).

I argue that this method should be treated with caution in dysphonia. It is questionable whether an acoustic model trained for normal speech recognition can be used to distinguish dysphonic speech at all. The training of an Automatic Speech Recognition system requires large amounts of training data, recorded from speakers who have different regional accents, voice characteristics, education backgrounds and genders etc. This data might include different accents and dialects, voices of smokers, young or elderly people. Hence, dysphonic speech, is possibly present, in larger numbers, in the database.The general goal of speech recognition is to recognize hoarse, nasal, sad, cheerful, old and young speech equally. I wanted to verify this hypothesis, whether using or adding phone posteriors might produce a better classification system for dysphonia than using acoustic features as input vector. To the best of my knowledge, this issue has never been evaluated for dysphonic speech.

I created the input vector from posterior probabilities of phones as described in Section 5.2.2, thus 21 phone posterior features were used as input, then the input vector was fed into a dysphonia classifier (DNN) and compared to the result presented in section 6.2.1. Results show that using acoustic features as the input vector of the classifier outperforms the ASR posterior features using DNN as a classifier. An accuracy of 88% was reached when acoustic features were used as input and 60% when ASR posterior features were used as input vector.

Since the ASR posterior features fall short behind the results obtained by the acoustic features, I examined whether the combination of the two input vectors (called "joint feature vector") increases the result of the classification accuracy. Results of the classifications are shown in Table 17. When the joint feature vector was used at the input of the neural network, the classification accuracy increased to 89%. While this is better than just using acoustic features, there is no significant impact of using ASR posterior probability values.

The confusion matrix of the DNN with the joint vector feature input can be seen in Table 18. The class precision of the HC class is 86.98%, while the precision of the Dys class is 89.92%. The recall for class HC is 86.53% and 90.27% for class Dys.

The class recall of Dys, when using the joint feature vector, is 90.27%. Which is higher than the 81.32% achieved when using only acoustic features. However, when comparing the joint feature vector's result with the case when acoustic features where used with FFS and SVM with linear kernel (presented in Table 15) the increase in the recall of Dys is not significant.

If false negatives and false positives have similar costs and chi-square test are performed, there is no significant difference between the classifications with accuracy of 85%, 88% and

Table 17: Two-class classification results between HC and Dys using DNN and comparing input vectors.

Input vector	**FFS**	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
Acoustic features	No	49	DNN	dropout	0.25	88%
ASR posterior features	No	21	DNN	dropout	0.25	60%
Joint features	No	70	DNN	dropout	0.25	**89%**

Table 18: Confusion matrix using a Fully-Connected Deep Neural Network with the joint features vector.

	true HC	**true Dys**	class precision
pred. HC	167	25	86.98%
pred. Dys	26	232	89.92%
class recall	86.53%	90.27%	

89%. The p-value between the acoustic features with SVM and the joint feature vector with DNN is 0.07, while between acoustic features with DNN and the joint feature vector with DNN is 0.91. Based on these, it can be concluded that it is not worthwhile to calculate ASR phone posterior, as it has no significant impact, but it can greatly complicate and slow down the current proposed system.

I calculated the frame level phone error rate (PER) to verify why the ASR posterior probability features failed to improve the classification accuracy. The DNN soft-max layer generates a 10 ms frame-level phone label that can be obtained by identifying the phone with the highest posterior. If the phone label matches with the reference given by the automatic phone segmentator's forced alignment (that was followed by manual corrections), this frame is said to be a matched frame. If the two 10 ms frames do not match, then the phone error rate (PER) is computed as the ratio of total number mismatched frames to the total number of frames in the reference (see equation 37). The mean PER on recordings containing healthy people's voice is 0.46, with 0.08 standard deviation, while it is 0.49 with 0.14 standard deviation on recordings containing dysphonic speech.

As mentioned in Section 6.1.2 four specialists were asked to evaluate the voice recordings

of the Selected Dysphonic and Healthy Database, thus a total of 148 recordings were evaluated by four specialists, the rest were evaluated by only one specialist. Since the internal consistency of the four specialists was high, I considered the dysphonia severity evaluation of the specialist, who gave the diagnosis, as true labels for the recordings. I calculated the Pearson correlation between the PER and the severity of dysphonia (H score from the RBH scale). The coefficient resulted in 0.2, indicating "weak" correlation. This further explains why the posterior probability features do not necessarily provide impact to the classification accuracy. I also calculated the Pearson correlations between the ASR posterior features and the H severity score (shown in Figures 13 and 14). In the figures ASR posterior features are abbreviated to 'post'.

As the figures show, the Pearson correlations vary in both cases from "very weak" to "moderate" correlations. When joining ASR posterior features with acoustic features at the input of the classifier, the ASR posterior features do not seem to provide additional relevant information.

The distributions of a specific phone's posteriors for the four severity categories (H) were calculated. In my first approach, I calculated the highest posterior (where the phoneme 'wins its frame') of the frames across the four severity categories. In the second approach, I calculated all the posteriors where the specific phone appeared. The results are shown in Figure 15 and 16 for phone [E] and Figure 17 and 18 for phone [h]. It shows that different severity categories do not separate well by the phone posteriors. Other phones have similar trends.

My obtained results do not necessarily contradict those described in [154]. In the work of Lee and his collegues, posterior features were examined in people with voice disorders. However, many voice disorders (e.g., tongue root tumours) are not associated with hoarseness. The RBH scale is primarily focused on hoarseness. Article [154] does not discuss the scale on which the severity of a voice disorder was interpreted, or the basis of what audible auditory characteristics were used to determine the severity of a voice disorder. This is an issue. On the RBH scale, the voice of a patient with a tongue tumour, cold or GERD may easily take on a value of R0B0H0, as these people are usually not hoarse, but there is an audibly severe change in their voice formation. In my work I consider patients sick if the H severity score is at least of value 1.

If we consider that dysphonia is quite common, and that the ASR model training corpora is not controlled w.r.t. hoarseness, then we may argue that ASR train sets, most likely,

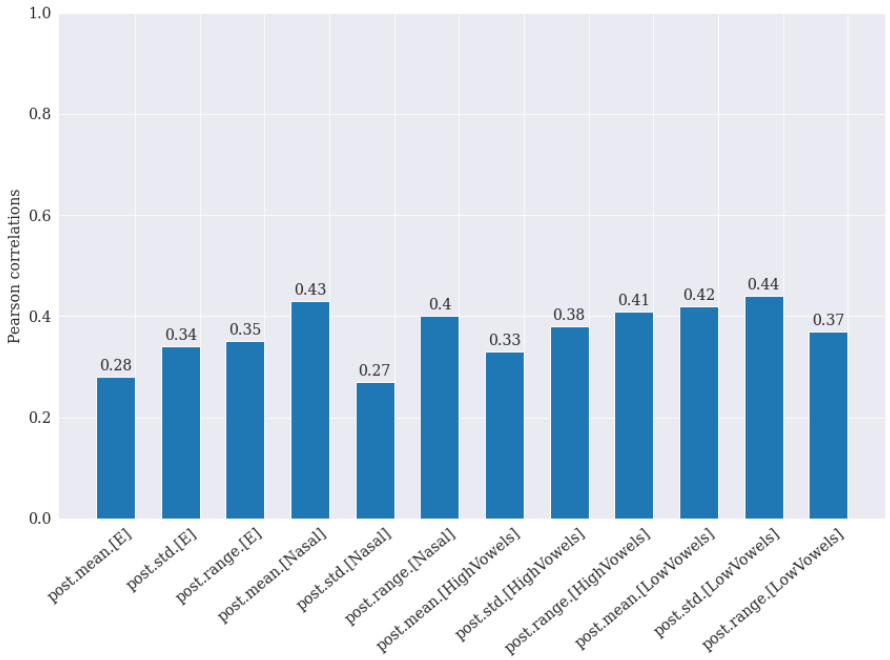

Figure 13: Pearson correlation between ASR posterior features and H.

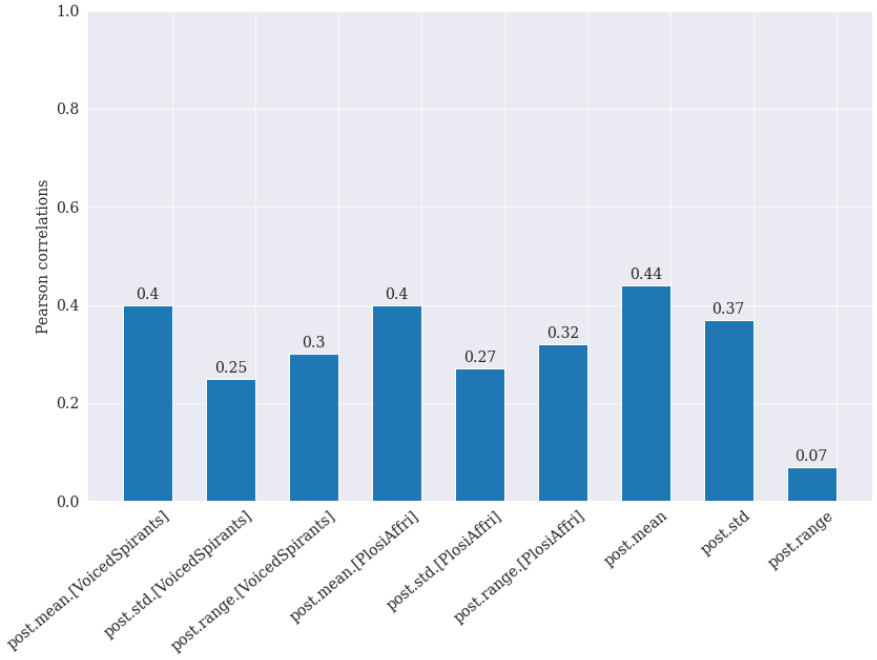

Figure 14: Pearson correlation between ASR posterior features and H.

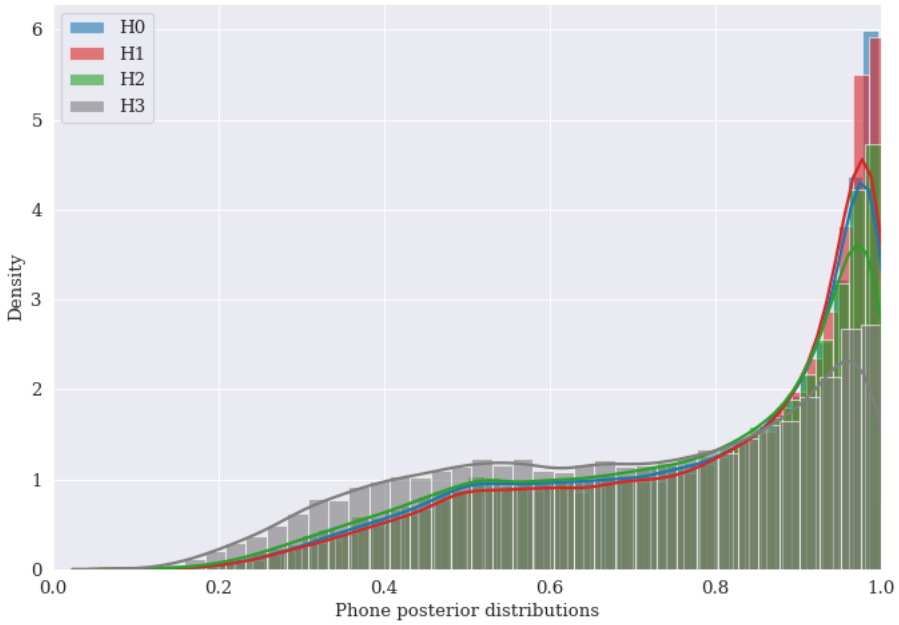

Figure 15: Phone posterior distributions of highest probability [E] phones.

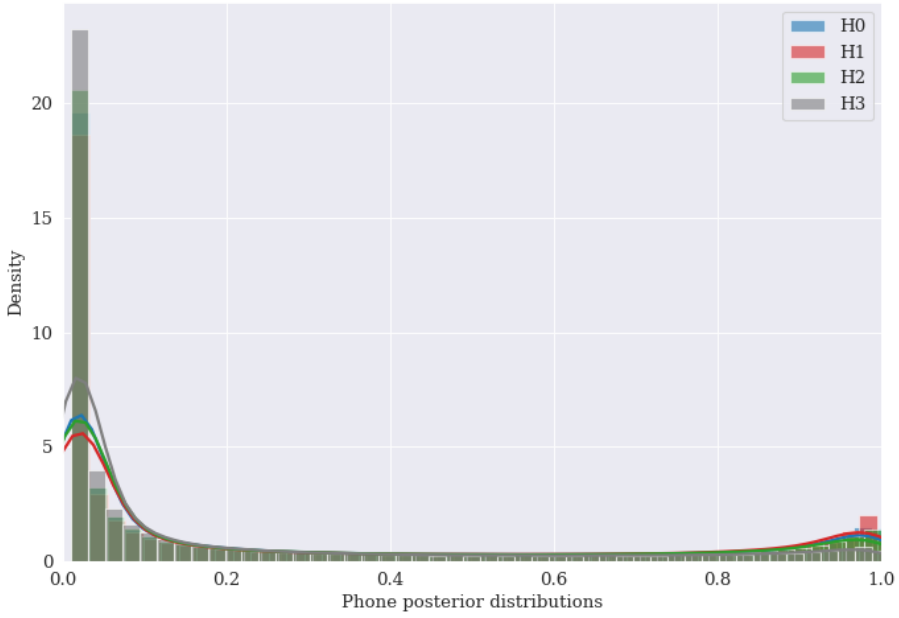

Figure 16: Phone posterior distributions of all [E] phones.

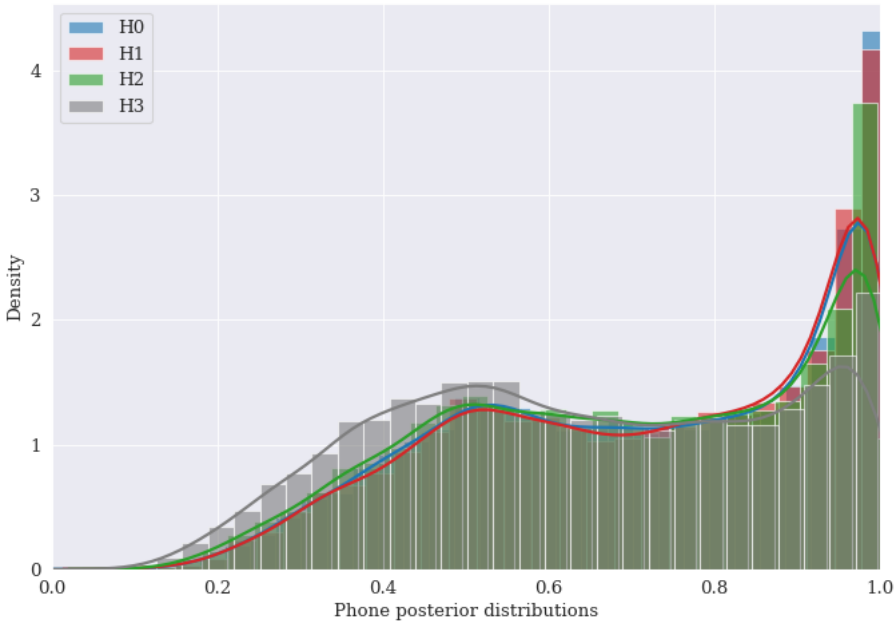

Figure 17: Phone posterior distributions of highest probability [h] phones.

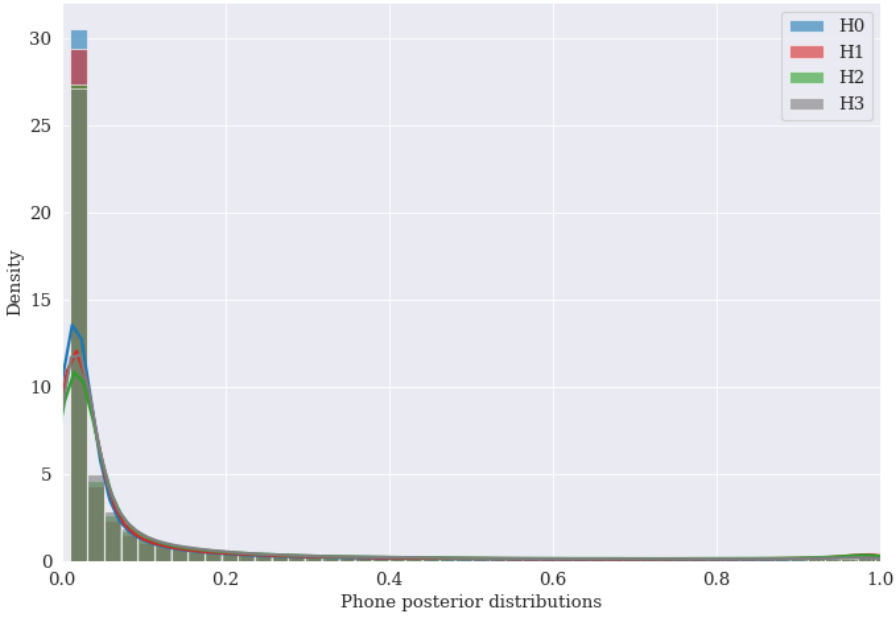

Figure 18: Phone posterior distributions of all [h] phones.

contain sufficient dysphonic samples to make the resulting models more tolerant to such kinds of deviation in voice characteristics. Therefore, whereas ASR posteriors can be a useful feature to address detection and classification of less frequent voice disorders – which in contrast have higher impact on voice quality than the hoarseness primarily associated with dysphonia – they are not ideal for the detection and classification of the severity of dysphonia.

From these results the following theses can be formulated.

Thesis II. B. [C2, C9] *I have shown, that using ASR phone posterior derived features, that were trained for general ASR purpose, is less effective in the automatic classification of healthy and dysphonic voices, than using the acoustic feature set directly. Deeper analyses showed weak relation between phone posterior distributions and dysphonia severity scores.*

Thesis II. C. [C2, C9] *I have shown that adding ASR phone posterior derived features to the acoustic features does not significantly improve the automatic classification accuracy of healthy and dysphonic voices.*

6.3 The automatic classification of functional and organic dysphonia

As mentioned in the Introduction (Section 1) dysphonia can be classified as either an organic or a functional disorder of the larynx.

According to Barth [155] and Stern [156] we are talking about a functional phonation disorder, if the diagnostic tools available to us do not detect organic lesions. The voice organs are healthy, yet the interplay of the temporal and dynamic systems of the factors necessary for voice production are disturbed. Weiss states that the "functional" indicator is temporary and valid only until the means of science can reveal the real organ causes of the illness [157]. Gundermann argues against this view and states it is not appropriate to use the term "functional" instead of "lack of organic" [158]. Organ abnormality can be the starting point of a functional disorder and vice versa, a functional disorder can lead to organic alteration. From the literature presented above, the two categories do not seem to be always mutually exclusive.

It is an interesting question whether it is possible to automatically separate functional from organic dysphonia. If functional dysphonia were determined with high probability with the help of a diagnosis support system, the patient would be directed to a phoniatrist or speech therapist. If the system detected organic dysphonia, the patient would be directed to an otolaryngologists or oncologist. This would save a lot of time and would lead the patient to care as soon as possible.

There are disputes in the definition and separation of FD and OD, and that the two categories may not be always mutually exclusive. It is natural that the two groups could be better classified on a database where the distributions of the severity of hoarseness for the two groups are statistically different, for example if the OD group has a statistically significant higher degree of severity than the FD group. In this way, the classifier may divide the severity of hoarseness (and not the disease types) into two groups: low and high. What we really want to achieve instead is to classify the two disease types in two. To investigate this phenomenon, the Filtered Dysphonic Database was created such a way that the distribution of the severity of hoarseness was not significantly different in the OD and FD groups.

In this Section I make an attempt to automatically separate functional from organic dysphonia. Although in subsection 6.2.1 I found that higher accuracy can be achieved using DNN, since the datasets I used in this experience are limited, I used SVM as a classifier. In

the present study two datasets were used, the Initial Dysphonic Database and the Filtered Dysphonic Database.

Initial Dysphonic Database

The Initial database contains all the 257 recordings form the database presented in Table 2 (described in Section 5.1.1) that were collected from patients suffering from dysphonia. From the Initial database 175 from patients were suffering from organic and 82 suffering from functional dysphonia. The database contains 156 females (95 with OD and 61 with FD) and 101 males (80 with OD and 21 with FD) recordings. The mean hoarseness score given for the OD group was 2 with 0.8 standard deviation (mean: 1.9, std.: 0.8 for females, mean: 2, std.: 0.9 for males), while for the FD group the mean was 1.4 with 0.7 standard deviation (mean: 1.3, std.: 0.7 for females, mean: 1.5, std.: 0.8 for males). The description of the Initial database is shown in Table 19.

Table 19: The Initial Dysphonic Database

	Number of female recordings	**Number of male recordings**	**H severity**	**Female H severity**	**Male H severity**
OD	95	80	2 (±0.8)	1.9 (±0.8)	2 (±0.9)
FD	61	21	1.4 (±0.7)	1.3 (±0.7)	1.5 (±0.8)

Filtered Dysphonic Database

The Filtered Dysphonic Database is the filtered version of the Initial Dysphonic Database described above. Filtered Dysphonic Database was created is such a way that the distribution of sexes and hoarseness levels in the OD and FD groups are equal. The Filtered Dysphonic Database contains a total number of 164 recordings, 82 from patients suffering from organic and 82 suffering from functional dysphonia. The database contains 122 females (61 with OD and 61 with FD) and 42 males (21 with OD and 21 with FD) recordings. The mean hoarseness score (H parameter from the RBH subjective scale) given for the OD group was 1.5 with 0.7 standard deviation, while for the FD group the mean was 1.4 with 0.7 standard deviation. The description of the Filtered Dysphonic Database is shown in Table 20.

In the experiment SVM with FFS was used with LOOCV data split technique. More on the SVM classifier and FFS algorithm in Section 5.4 and 5.4.6. As input, the 49 feature set was used as described in Section 5.2.1.

Table 20: The Filtered Dysphonic Database

	Number of female recordings	**Number of male recordings**	**H severity**	**Female H severity**	**Male H severity**
OD	61	21	1.5 (±0.7)	1.5 (±0.6)	1.5 (±0.8)
FD	61	21	1.4 (±0.7)	1.3 (±0.7)	1.5 (±0.8)

Classification using the Initial Dysphonic Database

Table 21 shows the classification results when the Initial Dysphonic Database was used. The highest accuracy was 76% using FFS parameter selection and SVM with linear kernel. The FFS algorithm selected 16 acoustic features, including the std and range of mfcc01, mean of HNR, std of shimmer, the std of SPI on nasals, the range of SPI on low vowels, the std and range of SPI on plosives, the mean, std and range of IMF entropy on nasals, the range of IMF entropy on high vowels, the mean, std and range of IMF entropy on low vowels and the std of IMF entropy on voiced spirants. The scenarios where FFS was used outperformed the scenarios where the original 49 acoustic feature set was used.

Table 21: Two-class classification results between OD and FD on the Initial Dysphonic Database.

FFS	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
No	49	SVM	linear kernel	C = 0.03125	66%
Yes	16	SVM	linear kernel	C = 1	**76%**
No	49	SVM	rbf kernel	C = 8; $\gamma = 0.5$	70%
Yes	5	SVM	rbf kernel	C = 0.25; $\gamma = 8$	70%

The confusion matrix of the classification can be seen in Table 22. As the table suggests, the class precision of the OD class is high (89%), but the precision of the FD class is much lower (59%). This means that the number of cases where the two classes were predicted correctly is not balanced. The recall for class FD was 80% and 74% for class OD.

Although the results are promising, the two classes differ in the severity of the hoarseness and the classifier may divide the severity of hoarseness (and not the disease types) into two groups: low and high. What we really want to achieve instead is to classify the two disease types in two. To investigate this phenomenon, the Filtered Dysphonic Database was created

Table 22: Confusion matrix using FFS and SVM with linear kernel on the Initial Dysphonic Database.

	true FD	**true OD**	class precision
pred. FD	61	27	69.32%
pred. OD	21	55	72.37%
class recall	74.39%	67.07%	

and statistical analysis was made.

Statistical analysis of the two databases

Since the severity scores are ordinal, the Mann-Whitney U test was used to check the statistical difference. Section 5.3 describes the Mann-Whitney U test in more detail. Significance level of 95% ($\alpha = 0.05$) was used.

Statistical analyses on the Initial Dysphonic Database

As Table 19 suggests, the mean hoarseness score given for the OD group was 2 with 0.8 standard deviation, while for the FD group the mean was 1.4 with 0.7 standard deviation. Using the Mann-Whitney U test the calculated p-value equals 0.000 for the total sample, 0.000 in case of females and 0.008 in case of males. Since p-value is less than α ($\alpha = 0.05$) in all three cases, the null hypothesis is rejected (the null hypothesis states that the two distributions are the same). The distribution of hoarseness of the OD group is considered not to be equal to the distribution hoarseness of the FD group. In other words, the difference between the distributions of hoarseness of the OD and FD groups is big enough to be statistically significant, this is true for females and males separately.

Statistical analyses on the Filtered Dysphonic Database

The Filtered Dysphonic Database is constructed with the purpose that it should not have significant difference in the distribution of the severity in the OD and FD dataset. As Table 20 shows, the mean hoarseness score is more balanced. When performing the Mann-Whitney U test, the calculated p-value equals to 0.650 for the total sample, 0.388 in case of females, 0.650 in case of males (p-value $> \alpha$). The distribution of hoarseness of the OD group is considered to be the same to the distribution of hoarseness of the FD group. In other words, the difference between the distributions of hoarseness of the OD and FD populations is not big enough to be statistically significant. Thus, the Filtered Dysphonic Database is suitable for further classification investigation as it excludes the possibility that the classifier is classifying the degree of hoarseness.

Classification using the Filtered Dysphonic Database

Table 23 shows the classification results in case when the Filtered Dysphonic Database was used. The highest accuracy was 71% using SVM with linear kernel. The FFS algorithm selected 5 acoustic features: the std of SPI on nasals, the std and range of IMF entropy measured on high vowels and the mean and std of IMF entropy on spirants. The scenarios where FFS was used outperformed the scenarios where the original 49 acoustic feature set was used in case of SVM with linear kernel, but not with rbf kernel.

Table 23: Two-class classification results between OD and FD on the Filtered Dysphonic Database.

FFS	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
No	49	SVM	linear kernel	C = 0.125	66%
Yes	5	SVM	linear kernel	C = 1	**71%**
No	49	SVM	rbf kernel	C = 128; $\gamma = 0.0005$	67%
Yes	10	SVM	rbf kernel	C = 2; $\gamma = 0.00781$	66%

The confusion matrix of the 71% classification result can be seen in Table 24. The class precision of the OD class is 72.37% and 69.32% in case of FD. The recall for class FD was 74.39% and 67.07% for class OD. This classification accuracy value is more reliable, than the classification on the dysphonic recordings from the database presented in Section 5.1.1, since it is unaffected by the difference in severity of hoarseness between the two groups.

Table 24: Confusion matrix using FFS and SVM with linear kernel on the Filtered Dysphonic Database.

	true FD	**true OD**	class precision
pred. FD	61	27	69.32%
pred. OD	21	55	72.37%
class recall	74.39%	67.07%	

The results clearly indicate that the separation between the two diseases can be done. I have mentioned that there are disputes in the definition and separation of FD and OD, and that the two categories may not be always mutually exclusive [155, 156, 157, 158]. Based on our experiment, it is clear that some acoustic parameters of speech change in case of FD,

and the automatic separation from OD is possible. This means that the acoustic parameters in FD change differently than in case of OD, so their separate group treatment is justified.

Thesis III. A. [C3] *I showed that the automatic separation between organic and functional dysphonia based on acoustic features is possible with 71% accuracy using an SVM with linear kernel on the Hungarian Filtered Dysphonic Database.*

6.4 The automatic classification of the voices of children with dysphonia

In Section 4.3 I presented a brief summary of some of the previous studies regarding the analysis of pathological children's voice. In my previous work [87] samples from healthy children and adult voices were compared giving a clear conclusion that differences exist in the examined acoustical parameters even between the two groups. But it is necessary to carry out the investigations separately on children's voices as well, we cannot use adult voices to draw any conclusions regarding children's voices.

Researchers mostly work with small sample sizes because it is difficult to collect recordings from children. In Section 5.1.2 I presented the Dysphonic and Healthy Child Speech Database which was built during the years of my research. It was essential to create a well-structured speech database containing children's speech samples, both from healthy children and children suffering from dysphonia. The database contains 59 recordings: 25 voices from children with dysphonia and 34 healthy children. Three children from the dysphonic group had vocal nodes, the rest had functional dysphonia.

The goal of this section is to make an attempt to automatically distinguish healthy voices of children from ones with dysphonia using SVM. The acoustic parameters used in this experiment are presented in Section 5.2.1.

For the binary classification an SVM classifier was used with linear and radial basis function (rbf) kernel. First, all 103 features calculated were used as input, then the FFS algorithm was used to reduce the dimensionality of the input vector. Usually in the case of rbf kernel the hyperparameter C is set to the number of parameters, while γ is set to 1/number of parameters. During the examination I used grid search to tune the hyperparameters. Leave-one-out cross validation was used in all cases. Classification results are summarized in Table 25.

As the table shows that the highest accuracy of 93% was reached using linear and rbf kernel. The features selection algorithm reduced the input dimensionality to 8 acoustic features, while achieving higher accuracy than in the case when all the starting features were used.

The confusion matrix can be seen in Table 26 when FFS and linear kernel was used. The class precision of the Healthy class is 94% and 92% in case of Dysphonia. The recall for the HC class was 94% and 92% for Dys.

Table 25: Two-class classification results on the Dysphonic and Healthy Child Speech Database.

FFS	**Number of features**	**Classifier and configuration**		**Hyper-parameters**	**LOOCV accuracy**
No	103	SVM	linear kernel	C = 1	88%
No	103	SVM	rbf kernel	C = 124; $\gamma = 0.008$	86%
Yes	8	SVM	linear kernel	C= 1	**93%**
Yes	8	SVM	rbf kernel	C = 10; $\gamma = 0.1$	**93%**

A successful classification should have a symmetric confusion matrix if the weights of the mismatch of the two classes are even. Otherwise, an asymmetric confusion matrix might indicate a biased classifier. The confusion matrix presented with 93% accuracy is as symmetric. However, as I mentioned earlier, confusion matrix is not symmetric in a medical system. Predicting a healthy child as dysphonic is less bad than predicting a child with dysphonia as healthy.

Furthermore, in my research it is essential that the number of true Dys cases misclassified as HC should be minimized. This happens only twice, resulting in a high 92% recall of class Dys.

Table 26: Confusion matrix using FFS and SVM with linear kernel.

	true HC	**true Dys**	class precision
pred. HC	32	2	94%
pred. Dys	2	23	92%
class recall	94%	92%	

The trend however is clear and promising; the automatic separation of healthy from pathological voices in the case of children is possible. We can conclude that using input vectors built by acoustic features have great power to distinguish healthy from dysphonic voices of children. Based on this result, it seems that dysphonia can be better screened at an early age, but much more data need to be collected to make such statements. This research can be a reference point in the classification of the voices of healthy children and voices of children with dysphonia using continuous speech.

Thesis IV. A. [J2] *I showed that the automatic separation of the voices of healthy children and children with dysphonia is possible. A classification accuracy of 93% can be achieved using SVM with linear or rbf kernel on the Dysphonic and Healthy Child Speech Database.*

7 Applicability of my results

The results demonstrate that developing a diagnosis support system which can differentiate dysphonic speech from healthy one is practically feasible. It is important to note, that while the system could be used for pre-screening, giving an exact diagnosis remains the responsibility of the physician.

The system proposed for adults comprises several steps: the speech recordings of the patients are arranged into speech databases (Dysphonic and Healthy Adult Speech Database). The recordings are normalized and segmented on phone level. After selecting the phones to be analysed, acoustic features are extracted and arranged into a feature vector. The feature vector is given to a classifier to perform the binary classification (healthy or unhealthy) in possession of prior knowledge. If the recording is classified as healthy the process stops. If it is classified as unhealthy this practical diagnosis support system would recognize the type of dysphonia, namely: functional or organic dysphonia, whilst performing the estimation of the severity of dysphonia based on a regression module.

Prior knowledge is gained by the procession of a carefully built speech database and optimal classification and regression models described in Sections 5.1.1, 6.1, 6.2, 6.3 and 6.4.

The class (healthy / unhealthy) or the severity of dysphonia is unknown for new speech samples. The preprocessing of the speech record is the same and after the acoustic features are measured on phone level a testing feature vector is constructed that enters a comparative unit, thus the classifications or regression are performed. This process is summarized in Fig 19.

If functional dysphonia can be determined with high probability, with the help of a diagnosis support system, the patient would be directed to a phoniatrist or speech therapist. If the system detected organic dysphonia, the patient would be directed to an otolaryngologists or oncologist. This would save a lot of time and would lead the patient to care as soon as possible. The end system proposed in this study can help young physicians or general practitioners filter out patients with dysphonia more efficiently and determine the severity of dysphonia automatically.

A diagnostic support system for the early recognition of dysphonia in the voices of children would follow the same logic described above, with the difference that the separation of organic from functional causes of dysphonia is not yet possible. Since the classification results in case of children's voice are promising, collecting further speech records to generalize

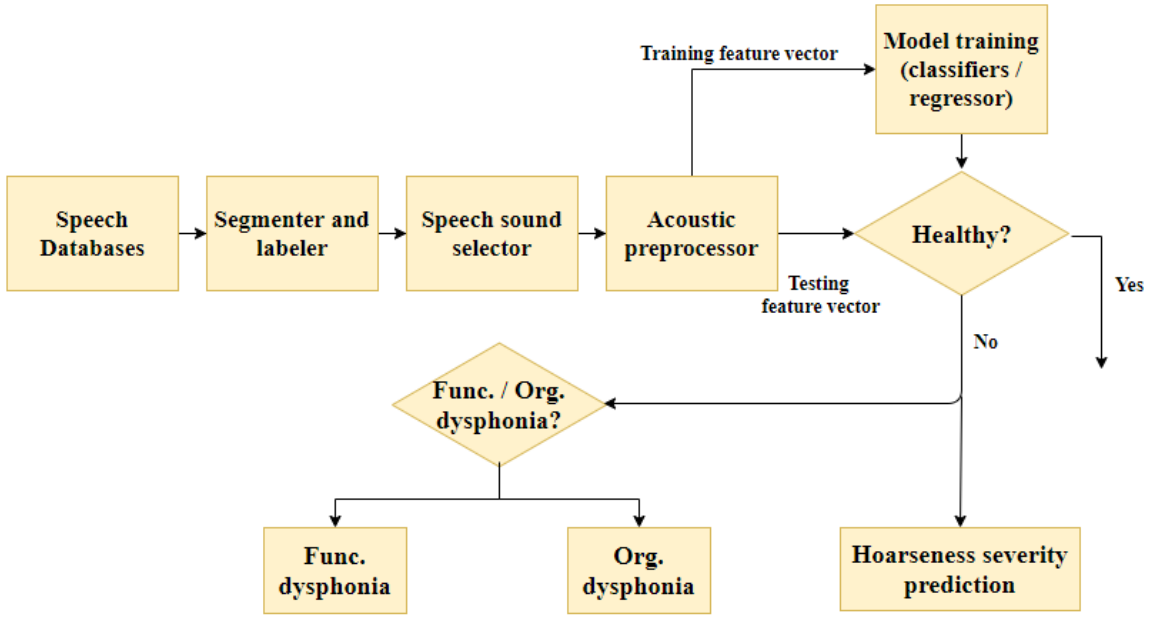

Figure 19: Proposed framework of a practical diagnosis support system for adults.

the classification model on a larger dataset is advised. In the long term it is worth developing a tool for the automatic detection of dysphonic voices among children. Mobile devices are suitable for implementing this method and using it in practice. Mobile health applications are usually designed for smart-phones or tablets, on some occasions smart-watches. They allow users to access information when and where they need it; reducing the time wasted with searching for specific data. These devices are cheap, easy-to-use and lightweight. Voice samples, metadata, acoustic feature values and the classifier output can be collected and uploaded to a cloud server. In this way, we can monitor the quality of the children's voice over the long term. The goal is to build a screening system that can be used by pre-school workers. If a child with dysphonic voice can be filtered in time, they will have a better chance of getting a professional help from an ear, nose and throat (ENT) specialist or a speech therapist.

8 Summary of my theses

In my Thesis I examined the effect of dysphonia on speech with the use of acoustic-phonetic features. With the help of these acoustic features, I examined the possibilities of automatic separation of a healthy voice from dysphonia, as well as the possibility of further automatic separation of the dysphonic voices according to the type of dysphonia. I also examined the possibilities of the estimation of the severity of dysphonia.

I have shown the importance of using feature selection, namely that its use results in a significant improvement in the recognition of dysphonia and in the estimation of the severity of dysphonia.

I proved by correlation analysis that some of the derived acoustic-phonetic features of speech change significantly under the influence of dysphonia, they significantly correlate with the severity of the hoarseness. I examined the auditory judgement of specialists to quantify the severity of dysphonia. I proved that there is a high degree of consistency among specialists when determining the severity of dysphonia using a popular subjective auditory scale (RBH). In order to examine the subjective nature of the RBH scale, k-means cluster analysis was done. I compared four cluster models, each of them labelled by one specialist's judgement. I showed that the severity of dysphonia can be clustered into four classes. Then, using the mean RBH perceptual evaluation of specialists as target I created regression models for the automatic assessment of the severity of dysphonia.

I showed that the binary classification of dysphonic and healthy voices is possible for Hungarian as well. I tried out and compared different classifiers such as SVMs and a Fully-Connected Deep Neural Network, as well as different input vectors for the classifier, such as acoustic features, ASR posterior probability features and the combination of the two. I concluded that it is not worthwhile to calculate ASR phone posterior for the classification of healthy and dysphonic voices, as it has no significant impact on the proposed system, but it can greatly complicate and slow it down.

I made a successful attempt to automatically separate organic and functional dysphonia based on acoustic features. There are disputes in the medical definition and separation of organic and functional dysphonia, but my results show that their automatic separation is possible and their treatment as separate groups is justified.

Lastly, I showed that the automatic classification of healthy speech and dysphonic speech is possible for children's speech as well. With the data collected so far it seems that dysphonia

can be better screened at an early stage, but we need to collect much more data to support this assertion.

In the following, I list my theses presented in my work as well as a brief summary of them.

Thesis I. A. [C4] *I showed that jitter(ddp), shimmer(dda), Harmonics-to-Noise Ratio (HNR), mfcc01, Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based IMF entropy frequency band ratios measured at specific phones show significant correlation at the 0.01 level with the severity of dysphonia when measured on the Hungarian Dysphonic and Healthy Adult Speech Database.*

In my first theses group, I examined the possibilities of automatically estimation of the severity of dysphonia.

Thesis I. A. states that dysphonia has an effect on the values of the acoustic features discussed. I verified my thesis statistically with the help of the Hungarian Dysphonic and Healthy Adult Speech Database, using speech samples from healthy people and from people suffering from dysphonia. A further consequence of the thesis is that the more severe the condition of dysphonia is, the more the values of the acoustic features differ from the values measured in the voices of healthy people.

Thesis I. B. [C4, J4] *I showed that when clustering the data, with the selected acoustic features using k-means clustering, the found clusters correlate well with the severity of dysphonia. A 0.59 Pearson correlation was achieved between the cluster defined values and the mean of the four specialists' ratings.*

In this thesis I examined the subjective nature of the dysphonia severity judgements of four highly educated specialists. Using k-means clustering algorithm the observations were classified into clusters, then clusters where assigned to the severity scores and four cluster models were compared. I calculated the Pearson correlation between the cluster defined severity scores and the individual specialists' ratings and I also evaluated this using the mean RBH perceptual evaluation of the four specialists. From this experiment I could conclude that the acoustic feature set used for clustering is suitable for modelling the individual assessments of dysphonia severity, and that the cluster defined severity correlate well with the severity values given by the specialists. The highest correlation value was measured between the cluster defined severity scores and the mean of the ratings. The mean of the four specialists' rating is the true label I use for the regression analyses.

Thesis I. C. [C4, J4] *I showed that an automatic estimation of the severity of dysphonia is possible using only eight acoustic features as input vector with a SVR with linear kernel reaching 0.85 Pearson correlation and 0.46 RMSE on the Selected Dysphonic and Healthy Database.*

This thesis states that the automatic estimation of the severity of dysphonia is not only possible, but it is possible with a low error. The consistency among the 4 specialists is of high degree and it can be considered as a theoretical goal we want to achieve. The regression model created greatly approximates the decision of the professionals.

Thesis II. A. [C2, C9] *I showed that the binary classification of dysphonic and healthy voices is possible for Hungarian. When applying a Fully-Connected Deep Neural Network, an accuracy of 88% can be achieved with LOOCV using acoustic features as input on the Hungarian Dysphonic and Healthy Adult Speech Database.*

In my second theses group, I examined the possibilities of binary classification of dysphonic and healthy speech using different machine learning approaches. Thesis II. A. states that the binary classification of dysphonic and healthy voices is possible for Hungarian. I verified the statement of my thesis with the help of the Hungarian Dysphonic and Healthy Adult Speech Database, by testing SVM-based classification models and a Fully-Connected Deep Neural Network. The highest accuracy of 88% was achieved with the later classification model using acoustic features as an input vector. The DNN as a classifier outperforms the SVM approaches with 3.53% relative accuracy, but it is also important to lower the risk that a person with dysphonia is miss-classified as healthy. In this sense using FFS and SVM with linear kernel seems to be a better approach.

Thesis II. B. [C2, C9] *I have shown, that using ASR phone posterior derived features, that were trained for general ASR purpose, is less effective in the automatic classification of healthy and dysphonic voices, than using the acoustic feature set directly. Deeper analyses showed weak relation between phone posterior distributions and dysphonia severity scores.*

In Thesis II. B. and Thesis II. C. I verified that using or adding phone posteriors might produce a better classification system for dysphonia than using acoustic features as input vector. I showed that the direct use of ASR posterior features falls short behind the results obtained by the acoustic features, as 60% accuracy was achieved when ASR posterior features were used as input vector and fed into the Fully-Connected Deep Neural Network.

Thesis II. C. [C2, C9] *I have shown that adding ASR phone posterior derived features to the acoustic features does not significantly improve the automatic classification accuracy of healthy and dysphonic voices.*

In this thesis I examined whether the combination of the two input vectors increases the result of the classification accuracy. While this approach yielded better accuracy than using only acoustic features, there was no significant impact of using ASR posterior probability values. Using the combined feature vector as the input of the neural network, the classification accuracy increased to 89%. If false negatives and false positives have similar costs and chi-square test are performed there is no significant difference between the classifications with accuracy of 85% (using acoustic features as input with FFS and SVM with linear kernel), 88% (using acoustic features as input of the DNN) and 89%.

One explanation why ASR posterior features did not help improve the classification result is that ASR model training corpora are not controlled w.r.t. hoarseness ASR train sets most likely contain sufficient dysphonic sample to make the recognizer robust to dysphonic voices. Therefore, whereas ASR posteriors can be a useful feature to address detection and classification of less frequent voice disorders they are not ideal for the detection of dysphonia.

Thesis III. A. [C3] *I showed that the automatic separation between organic and functional dysphonia based on acoustic features is possible with 71% accuracy using SVM with linear kernel on the Hungarian Filtered Dysphonic Database.*

In this single standing thesis I made a successful attempt to automatically classify dysphonic voices by the type of dysphonia. Dysphonia can be classified into two fundamental categories: functional and organic dysphonia. This classification is much criticized as the two categories may not be always mutually exclusive, organic abnormality can cause a functional disorder, a functional disorder can lead to organic alteration. The two dysphonia types may be different in the severity of voice quality, this could affect the classification attempt. Thus, it is important to construct a database where there is no significant difference in the distribution of the severity of dysphonia in the two groups. The Hungarian Filtered Dysphonic Database was constructed in such way and is suitable for classification investigation as it excludes the possibility that the classifier is classifying the degree of hoarseness instead of the type of dysphonia. The Filtered Dysphonic Database is a subset of the Hungarian Dysphonic and Healthy Adult Speech Database. In this thesis I showed that the separation of organic and functional dysphonia can be done.

Thesis IV. A. [J2] *I showed that the automatic separation of the voices of healthy children and children with dysphonia is possible. A classification accuracy of 93% can be achieved using SVM with linear or rbf kernel on the Dysphonic and Healthy Child Speech Database.*

In my final single standing thesis I addressed the problem of recognizing dysphonia in children's voice. It is necessary to carry out the investigations separately on children's voices as we cannot use adult voices to draw any conclusions regarding children's voices. The thesis states that the automatic separation of the voices of healthy children and children with dysphonia is possible. I used the Dysphonic and Healthy Child Speech Database in order to prove this assertion. The Dysphonic and Healthy Child Speech Database currently contains 59 recordings: 25 voices from children with dysphonia and 34 healthy children. Much more data is needed to obtain better and more general results, plus to avoid overfitting issues. However it is clear, that the classification can be done with high accuracy and high recall for the dysphonia class.

References

[1] R. J. Stachler, D. O. Francis, S. R. Schwartz, C. C. Damask, G. P. Digoy, H. J. Krouse, S. J. McCoy, D. R. Ouellette, R. R. Patel, C. C. W. Reavis *et al.*, "Clinical practice guideline: hoarseness (dysphonia)(update)," *Otolaryngology–Head and Neck Surgery*, vol. 158, no. 1_suppl, pp. S1–S42, 2018.

[2] S. M. Cohen, J. Kim, N. Roy, C. Asche, and M. Courey, "Prevalence and causes of dysphonia in a large treatment-seeking population," *The Laryngoscope*, vol. 122, no. 2, pp. 343–348, 2012.

[3] S. M. Cohen, "Self-reported impact of dysphonia in a primary care population: An epidemiological study," *The Laryngoscope*, vol. 120, no. 10, pp. 2022–2032, 2010.

[4] R. Reiter, T. K. Hoffmann, A. Pickhard, and S. Brosch, "Hoarseness—causes and treatments," *Deutsches Ärzteblatt International*, vol. 112, no. 19, p. 329, 2015.

[5] K. Jones, J. Sigmon, L. Hock, E. Nelson, M. Sullivan, and F. Ogren, "Prevalence and risk factors for voice problems among telemarketers," *Archives of Otolaryngology–Head & Neck Surgery*, vol. 128, no. 5, pp. 571–577, 2002.

[6] J. Long, H. N. Williford, M. S. Olson, and V. Wolfe, "Voice problems and risk factors among aerobics instructors," *Journal of Voice*, vol. 12, no. 2, pp. 197–207, 1998.

[7] E. Smith, H. L. Kirchner, M. Taylor, H. Hoffman, and J. H. Lemke, "Voice problems among teachers: differences by gender and teaching characteristics," *Journal of Voice*, vol. 12, no. 3, pp. 328–334, 1998.

[8] T. Davids, A. M. Klein, and M. M. Johns III, "Current dysphonia trends in patients over the age of 65: is vocal atrophy becoming more prevalent?" *The Laryngoscope*, vol. 122, no. 2, pp. 332–335, 2012.

[9] N. Bhattacharyya, "The prevalence of pediatric voice and swallowing problems in the united states," *The Laryngoscope*, vol. 125, no. 3, pp. 746–750, 2015.

[10] M. C. Duff, A. Proctor, and E. Yairi, "Prevalence of voice disorders in african american and european american preschoolers," *Journal of Voice*, vol. 18, no. 3, pp. 348–353, 2004.

[11] P. N. Carding, S. Roulstone, K. Northstone, A. S. Team *et al.*, "The prevalence of childhood dysphonia: a cross-sectional study," *Journal of Voice*, vol. 20, no. 4, pp. 623–630, 2006.

[12] E.-M. Silverman and C. H. Zimmer, "Incidence of chronic hoarseness among school-age children," *Journal of Speech and Hearing Disorders*, vol. 40, no. 2, pp. 211–215, 1975.

[13] M. Rosa and M. Behlau, "Mapping of vocal risk in amateur choir," *Journal of Voice*, vol. 31, no. 1, pp. 118–e1, 2017.

[14] J. Guss, B. Sadoughi, B. Benson, and L. Sulica, "Dysphonia in performers: toward a clinical definition of laryngology of the performing voice," *Journal of Voice*, vol. 28, no. 3, pp. 349–355, 2014.

[15] K. Verdolini and L. O. Ramig, "Occupational risks for voice problems," *Logopedics Phoniatrics Vocology*, vol. 26, no. 1, pp. 37–46, 2001.

[16] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of voice disorders in teachers and the general population," *Journal of Speech, Language, and Hearing Research*, 2004.

[17] E. Smith, J. Lemke, M. Taylor, H. L. Kirchner, and H. Hoffman, "Frequency of voice problems among teachers and other occupations," *Journal of voice*, vol. 12, no. 4, pp. 480–488, 1998.

[18] F. S. G. Fortes, R. Imamura, D. H. Tsuji, and L. U. Sennes, "Profile of voice professionals seen in a tertiary health center," *Brazilian journal of otorhinolaryngology*, vol. 73, no. 1, pp. 27–31, 2007.

[19] D. Isetti and T. Meyer, "Workplace productivity and voice disorders: A cognitive interviewing study on presenteeism in individuals with spasmodic dysphonia," *Journal of Voice*, vol. 28, no. 6, pp. 700–710, 2014.

[20] "American speech-language-hearing association - voice disorders," https://www.asha.org/practice-portal/clinical-topics/voice-disorders/, accessed: 2020-04-03.

[21] R. Jani, S. Jaana, L. Laura, and V. Jos, "Systematic review of the treatment of functional dysphonia and prevention of voice disorders," *Otolaryngology—Head and Neck Surgery*, vol. 138, no. 5, pp. 557–565, 2008.

[22] K. Strimbu and J. A. Tavel, "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010.

[23] G. Olaszy, *A magyar beszéd*. Akadémia Kiadó, 2010.

[24] M. M. Johns III, R. T. Sataloff, A. L. Merati, and C. A. Rosen, "Shortfalls of the american academy of otolaryngology–head and neck surgery's clinical practice guideline: Hoarseness (dysphonia)," *Otolaryngology-Head and Neck Surgery*, vol. 143, no. 2, pp. 175–177, 2010.

[25] N. Bhattacharyya, "The prevalence of voice problems among adults in the united states," *The Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, 2014.

[26] S. M. Cohen, W. D. Dupont, and M. S. Courey, "Quality-of-life impact of non-neoplastic voice disorders: a meta-analysis," *Annals of Otology, Rhinology & Laryngology*, vol. 115, no. 2, pp. 128–134, 2006.

[27] M. S. Benninger, A. S. Ahuja, G. Gardner, and C. Grywalski, "Assessing outcomes for dysphonic patients," *Journal of voice*, vol. 12, no. 4, pp. 540–550, 1998.

[28] L. O. Ramig and K. Verdolini, "Treatment efficacy: voice disorders," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, pp. S101–S116, 1998.

[29] M. Behlau, G. Madazio, and G. Oliveira, "Functional dysphonia: strategies to improve patient outcomes," *Patient related outcome measures*, vol. 6, p. 243, 2015.

[30] J. Hirschberg, T. Hacki, and K. Mészáros, "Phoniatrics and related professions. [*F*oniátria és *T*árstudományok.]," *ELTE Eötvös Kiadó, Budapest*, 2013.

[31] I. R. Titze and D. W. Martin, "Principles of voice production," 1998.

[32] T. Hacki, M. Moerman, and J. S. Rubin, "'malregulative' rather than 'functional'dysphonia: A new etiological terminology framework for phonation disorders—a position paper by the union of european phoniatricians (uep)," *Journal of Voice*, 2020.

[33] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith, "Voice disorders in the general population: prevalence, risk factors, and occupational impact," *The Laryngoscope*, vol. 115, no. 11, pp. 1988–1995, 2005.

[34] S. M. Coyle, B. D. Weinrich, and J. C. Stemple, "Shifts in relative prevalence of laryngeal pathology in a treatment-seeking population," *Journal of Voice*, vol. 15, no. 3, pp. 424–440, 2001.

[35] I. R. Titze, J. Lemke, and D. Montequin, "Populations in the us workforce who rely on voice as a primary tool of trade: a preliminary report," *Journal of voice*, vol. 11, no. 3, pp. 254–259, 1997.

[36] N. Roy, J. Kim, M. Courey, and S. M. Cohen, "Voice disorders in the elderly: a national database study," *The Laryngoscope*, vol. 126, no. 2, pp. 421–428, 2016.

[37] S. M. Cohen, J. Kim, N. Roy, C. Asche, and M. Courey, "Direct health care costs of laryngeal diseases and disorders," *The Laryngoscope*, vol. 122, no. 7, pp. 1582–1588, 2012.

[38] D. O. Francis, M. E. McKiever, C. G. Garrett, B. Jacobson, and D. F. Penson, "Assessment of patient experience with unilateral vocal fold immobility: a preliminary study," *Journal of Voice*, vol. 28, no. 5, pp. 636–643, 2014.

[39] N. Mirza, C. Ruiz, E. D. Baum, and J. P. Staab, "The prevalence of major psychiatric pathologies in patients with voice disorders," *Ear, nose & throat journal*, vol. 82, no. 10, pp. 808–814, 2003.

[40] A. Y. Chen, N. M. Schrag, M. Halpern, A. Stewart, and E. M. Ward, "Health insurance and stage at diagnosis of laryngeal cancer: does insurance type predict stage at diagnosis?" *Archives of Otolaryngology–Head & Neck Surgery*, vol. 133, no. 8, pp. 784–790, 2007.

[41] N. Roy, J. Stemple, R. M. Merrill, and L. Thomas, "Epidemiology of voice disorders in the elderly: preliminary findings," *The Laryngoscope*, vol. 117, no. 4, pp. 628–633, 2007.

[42] J. S. Golub, P.-H. Chen, K. J. Otto, E. Hapner, and M. M. Johns, "Prevalence of perceived dysphonia in a geriatric population," *Journal of the American Geriatrics Society*, vol. 54, no. 11, pp. 1736–1739, 2006.

[43] R. H. G. Martins, E. R. B. N. Pereira, C. B. Hidalgo, and E. L. M. Tavares, "Voice disorders in teachers. a review," *Journal of Voice*, vol. 28, no. 6, pp. 716–724, 2014.

[44] D. Hazlett, O. Duffy, and S. Moorhead, "Review of the impact of voice training on the vocal quality of professional voice users: implications for vocal health and recommendations for further research," *Journal of voice*, vol. 25, no. 2, pp. 181–191, 2011.

[45] R. E. de la Hoz, M. R. Shohet, L. A. Bienenfeld, A. A. Afilaka, S. M. Levin, and R. Herbert, "Vocal cord dysfunction in former world trade center (wtc) rescue and recovery workers and volunteers," *American journal of industrial medicine*, vol. 51, no. 3, pp. 161–165, 2008.

[46] E. E. Levendoski, A. Sundarrajan, and M. P. Sivasankar, "Reducing the negative vocal effects of superficial laryngeal dehydration with humidification," *Annals of Otology, Rhinology & Laryngology*, vol. 123, no. 7, pp. 475–481, 2014.

[47] M. Hirano, "*GRBAS* scale for evaluating the hoarse voice & frequency range of phonation," *Clinical examination of voice*, vol. 5, pp. 83–84, 1981.

[48] K. Omori, "Diagnosis of voice disorders," *JMAJ*, vol. 54, no. 4, pp. 248–253, 2011.

[49] M. Ptok, C. Schwemmle, C. Iven, M. Jessen, and T. Nawka, "On the auditory evaluation of voice quality," *HNO*, vol. 54, no. 10, pp. 793–802, 2006.

[50] J. Wendler, A. Rauhut, and H. Kruger, "Classification of voice qualities," *Journal of Phonetics*, vol. 14, no. 3-4, pp. 483–488, 1986.

[51] E. Seifert, "Stress and distress in non-organic voice disorder," *Swiss medical weekly*, vol. 135, no. 2728, 2005.

[52] F. L. Wuyts, M. S. D. Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. V. Lierde, J. Raes, and P. H. V. d. Heyning, "The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach," *Journal of speech, language, and hearing research*, vol. 43, no. 3, pp. 796–809, 2000.

[53] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 95–99.

[54] Y. Zhang and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies," *Journal of Voice*, vol. 22, no. 1, pp. 1–9, 2008.

[55] Z. Ali, M. Talha, and M. Alsulaiman, "A practical approach: Design and implementation of a healthcare software for screening of dysphonic patients," *IEEE Access*, vol. 5, pp. 5844–5857, 2017.

[56] J. P. Teixeira, P. O. Fernandes, and N. Alves, "Vocal acoustic analysis–classification of dysphonic voices with artificial neural networks," *Procedia computer science*, vol. 121, pp. 19–26, 2017.

[57] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. F. Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2017.

[58] V. Klára, I. Viktor, and M. Krisztina, "Voice disorder detection on the basis of continuous speech," in *5th European Conference of the International Federation for Medical and Biological Engineering*. Springer, 2011, pp. 86–89.

[59] V. Guedes, F. Teixeira, A. Oliveira, J. Fernandes, L. Silva, A. Junior, and J. P. Teixeira, "Transfer learning with audioset to voice pathologies identification in continuous speech," *Procedia Computer Science*, vol. 164, pp. 662–669, 2019.

[60] H. Cordeiro, C. Meneses, and J. Fonseca, "Continuous speech classification systems for voice pathologies identification," in *Doctoral Conference on Computing, Electrical and Industrial Systems*. Springer, 2015, pp. 217–224.

[61] S. Lahmiri, D. A. Dawson, and A. Shmuel, "Performance of machine learning methods in diagnosing parkinson's disease based on dysphonia measures," *Biomedical engineering letters*, vol. 8, no. 1, pp. 29–39, 2018.

[62] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[63] G. Gidaye, J. Nirmal, K. Ezzine, A. Shrivas, and M. Frikha, "Application of glottal flow descriptors for pathological voice diagnosis," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 205–222, 2020.

[64] Y. Liu, T. Lee, P. Ching, T. K. Law, and K. Y. Lee, "Acoustic assessment of disordered voice with continuous speech based on utterance-level *ASR* posterior features." in *Interspeech*, 2017, pp. 2680–2684.

[65] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "Convolutional neural networks for pathological voice detection," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1–4.

[66] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[67] A. Rueda and S. Krishnan, "Augmenting dysphonia voice using fourier-based synchrosqueezing transform for a *CNN* classifier," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6415–6419.

[68] 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*, vol. 31, pp. 156–164, 2017.

[69] F. Kazinczi, K. Mészáros, and K. Vicsi, "Automatic detection of voice disorders," in *International Conference on Statistical Language and Speech Processing*. Springer, 2015, pp. 143–152.

[70] G. Muhammad, M. Alsulaiman, A. Mahmood, and Z. Ali, "Automatic voice disorder classification using vowel formants," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.

[71] T. L. Eadie and C. R. Baylor, "The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice," *Journal of voice*, vol. 20, no. 4, pp. 527–544, 2006.

[72] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. De Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality:

combining continuous speech and sustained vowels," *Journal of voice*, vol. 24, no. 5, pp. 540–555, 2010.

[73] Y.-R. Chien, M. Borskỳ, and J. Guðnason, "Objective severity assessment from disordered voice using estimated glottal airflow." in *Interspeech*, 2017, pp. 304–308.

[74] V. Wolfe, R. Cornell, and J. Fitch, "Sentence/vowel correlation in the evaluation of dysphonia," *Journal of Voice*, vol. 9, no. 3, pp. 297–303, 1995.

[75] Y. Maryn and N. Roy, "Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity," *Jornal da Sociedade Brasileira de Fonoaudiologia*, vol. 24, no. 2, pp. 107–112, 2012.

[76] M. De Bodt, "A framework for voice assessment: The relation between subjective and objective parameters in the judgment of normal and pathological voice," *Antwerp, Belgium: University of Antwerp*, 1997.

[77] R. I. Zraick, K. Wendel, and L. Smith-Olinde, "The effect of speaking task on perceptual judgment of the severity of dysphonic voice," *Journal of Voice*, vol. 19, no. 4, pp. 574–581, 2005.

[78] S. Fujimura, T. Kojima, Y. Okanoue, K. Shoji, M. Inoue, K. Omori, and R. Hori, "Classification of voice disorders using a one-dimensional convolutional neural network," *Journal of Voice*, 2020.

[79] B. B. v. Latoszek, N. Ulozaitė-Stanienė, Y. Maryn, T. Petrauskas, and V. Uloza, "The influence of gender and age on the acoustic voice quality index and dysphonia severity index: A normative study," *Journal of Voice*, vol. 33, no. 3, pp. 340–345, 2019.

[80] M. G. Tulics and K. Vicsi, "The automatic assessment of the severity of dysphonia," *International Journal of Speech Technology*, vol. 22, no. 2, pp. 341–350, 2019.

[81] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the *GRBAS* scale," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 2478–2481.

[82] K. Szklanny and P. Wrzeciono, "Relation of *RBH* auditory-perceptual scale to acoustic and electroglottographic voice analysis in children with vocal nodules," *IEEE Access*, vol. 7, pp. 41 647–41 658, 2019.

[83] S. K. Swain, B. Nahak, L. Sahoo, S. Munjal, and M. C. Sahu, "Pediatric dysphonia-a review," *Indian J Child Health*, vol. 6, pp. 1–5, 2019.

[84] M. A. Kiliç, E. Okur, I. Yildirim, and S. Güzelsoy, "The prevalence of vocal fold nodules in school age children," *International Journal of Pediatric Otorhinolaryngology*, vol. 68, no. 4, pp. 409–412, 2004.

[85] R. K. Shah, G. H. Woodnorth, A. Glynn, and R. C. Nuss, "Pediatric vocal nodules: correlation with perceptual voice analysis," *International journal of pediatric otorhinolaryngology*, vol. 69, no. 7, pp. 903–909, 2005.

[86] R. K. Shah, S. H. Engel, and S. S. Choi, "Relationship between voice quality and vocal nodule size," *Otolaryngology–Head and Neck Surgery*, vol. 139, no. 5, pp. 723–726, 2008.

[87] M. G. Tulics, F. Kazinczi, and K. Vicsi, "Statistical analysis of acoustical parameters in the voice of children with juvenile dysphonia," in *International Conference on Speech and Computer*. Springer, 2016, pp. 667–674.

[88] M. L. Meredith, S. M. Theis, J. S. McMurray, Y. Zhang, and J. J. Jiang, "Describing pediatric dysphonia with nonlinear dynamic parameters," *International journal of pediatric otorhinolaryngology*, vol. 72, no. 12, pp. 1829–1836, 2008.

[89] K. E. Joseph E.Dohar, Amber D.Shaffer, "Pediatric dysphonia: It's not about the nodules," *International Journal of Pediatric Otorhinolaryngology*, vol. 125, no. 12, pp. 147–152, 2019.

[90] J. Coelho, D. Ramos, I. Monteiro, and A. D. Paiva, "Vocal nodules in school age children," *Revista de Logopedia, Foniatría y Audiología*, vol. 36, no. 3, pp. 103–108, 2016.

[91] G. K. Pebbili, J. Kidwai, and S. Shabnam, "Dysphonia severity index in typically developing indian children," *Journal of Voice*, vol. 31, no. 1, pp. 125–e1, 2017.

[92] D. D. Deliyski, H. S. Shaw, M. K. Evans, and R. Vesselinov, "Regression tree approach to studying factors influencing acoustic voice analysis," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 4, pp. 274–288, 2006.

[93] G. Kiss and K. Vicsi, "Mono-and multi-lingual depression prediction based on speech processing," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 919–935, 2017.

[94] V. Klára, "Sampa computer readable phonetic alphabet," 2008.

[95] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, vol. 454, no. 1971. The Royal Society, 1998, pp. 903–995.

[96] A. Tsanas, "Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms," *Models and analysis of vocal emissions for biomedical applications*, vol. 2, pp. 37–40, 2013.

[97] A. Zeiler, R. Faltermeier, I. R. Keck, A. M. Tomé, C. G. Puntonet, and E. W. Lang, "Empirical mode decomposition-an introduction," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–8.

[98] T. Andrea, "Alaphangjellemzők vizsgálata gyermekek beszédében," *Alknyelvdok9*, p. 62.

[99] I. Titze, "Principles of voice production (national center for voice and speech, iowa city, ia)," *Chap*, vol. 6, pp. 149–184, 2000.

[100] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[101] P. Dighe, A. Asaei, and H. Bourlard, "On quantifying the quality of acoustic models in hybrid *DNN-HMM ASR*," *Speech Communication*, 2020.

[102] P. Roach, S. Arnfield, W. Barry, J. Baltova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel *et al.*, "Babel: An eastern european multilanguage database," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1892–1893.

[103] K. Vicsi, A. Kocsor, C. Teleki, and L. Tóth, "Hungarian speech database for computer-using environments in offices," in *Proc. 2nd Hungarian Conf. on Computational Linguistics*, 2004, pp. 315–318.

[104] C. Teleki, V. Szabolcs, T. S. Levente, and V. Klára, "Development and evaluation of a hungarian broadcast news database," in *Forum Acusticum*, 2005.

[105] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[106] W. G. Cochran, "The $\chi 2$ test of goodness of fit," *The Annals of Mathematical Statistics*, pp. 315–345, 1952.

[107] D. Elek, *Biometria az orvosi gyakorlatban.* Medicina Könyvkiadó Zrt., 2011.

[108] B. Sándor, *Bevezetés a matematikai statisztikába.* Kossuth Egyetemi Kiadó, 1997.

[109] J. D. Evans, *Straightforward statistics for the behavioral sciences.* Thomson Brooks/Cole Publishing Co, 1996.

[110] J. M. Cortina, "What is coefficient alpha? an examination of theory and applications." *Journal of applied psychology*, vol. 78, no. 1, p. 98, 1993.

[111] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International journal of medical education*, vol. 2, p. 53, 2011.

[112] I. Trizano-Hermosilla and J. M. Alvarado, "Best alternatives to cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements," *Frontiers in psychology*, vol. 7, p. 769, 2016.

[113] H. Wold, "Encyclopedia of statistical sciences," *Partial least squares. Wiley, New York*, pp. 581–591, 1985.

[114] C. A. Bobak, P. J. Barr, and A. J. O'Malley, "Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales," *BMC medical research methodology*, vol. 18, no. 1, p. 93, 2018.

[115] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." *Psychological assessment*, vol. 6, no. 4, p. 284, 1994.

[116] G. Horváth, "Neurális hálózatok és műszaki alkalmazásaik," *Műszaki Kiadó*, 1998.

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

[118] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[119] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[120] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[121] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in neural information processing systems*, 1997, pp. 155–161.

[122] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[123] M. Székelyi and B. Ildiko, *Tulelokeszlet az SPSS-hez*. Typotex kiadó, 2002.

[124] S. Russell and P. Norvig, "Mesterséges intelligencia, modern megközelítésben, második, átdolgozott, bővített kiadás, budapest, panem kiadó, 2005," *Közvetlen link: http://project. mit. bme. hu/mi_almanach/books/aima/index*.

[125] Y. Chauvin and D. E. Rumelhart, *Backpropagation: theory, architectures, and applications*. Psychology press, 1995.

[126] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[127] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 421–436.

[128] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[129] M. Altrichter, G. Horváth, B. Pataki, G. Strausz, G. Takács, J. Valyon, B. Czétényi, I. T. Engedy, and K. Gáti, "Neurális hálózatok," 2006.

[130] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, p. 91, 2006.

[131] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[132] J. Wainer and G. Cawley, "Nested cross-validation when selecting classifiers is overzealous for most practical applications," *Journal of Machine Learning Researc*, 2017.

[133] L. Luo, L. Ye, M. Luo, D. Huang, H. Peng, and F. Yang, "Methods of forward feature selection based on the aggregation of classifiers generated by single attribute," *Computers in biology and medicine*, vol. 41, no. 7, pp. 435–441, 2011.

[134] S. D. Kiss, Gabor and K. Vicsi, "Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features," in *IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2013, pp. 579–582.

[135] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[136] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[137] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[138] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[139] I. Corp., "Ibm spss statistics for windows." [Online]. Available: https://www.ibm.com/analytics/spss-statistics-software

[140] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications.* CRC Press, 2016.

[141] D. Laszuk, "Python implementation of empirical mode decomposition algorithm," https://github.com/laszukdawid/PyEMD, 2017.

[142] I. Laaridh, W. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Interspeech*, 2017.

[143] T. Law, J. H. Kim, K. Y. Lee, E. C. Tang, J. H. Lam, A. C. van Hasselt, and M. C. Tong, "Comparison of rater's reliability on perceptual evaluation of different types of voice sample," *Journal of Voice*, vol. 26, no. 5, pp. 666–e13, 2012.

[144] N. Adiga, C. Vikram, K. Pullela, and S. M. Prasanna, "Zero frequency filter based analysis of voice disorders." in *Interspeech*, 2017, pp. 1824–1828.

[145] M. Markaki, Y. Stylianou, J. D. Arias-Londoño, and J. I. Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2010, pp. 5162–5165.

[146] P. Harar, J. B. Alonso-Hernandezy, 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).* IEEE, 2017, pp. 1–4.

[147] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, 2018.

[148] H. Wu, J. J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks." in *Interspeech*, vol. 2018, 2018.

[149] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[150] L. Salhi, M. Talbi, and A. Cherif, "Voice disorders identification using hybrid approach: Wavelet analysis and multilayer neural networks," *World Academy of Science, Engineering and Technology*, vol. 45, no. 21, pp. 330–339, 2008.

[151] V. Srinivasan, V. Ramalingam, and P. Arulmozhi, "Artificial neural network based pathological voice classification using mfcc features," *International Journal of Science, Environment and Technology*, vol. 3, no. 1, pp. 291–302, 2014.

[152] C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech & Language*, vol. 4, no. 2, pp. 127–165, 1990.

[153] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[154] T. Lee, Y. Liu, Y. T. Yeung, T. K. Law, and K. Y. Lee, "Predicting severity of voice disorder from *DNN-HMM* acoustic posteriors." in *Interspeech*, 2016, pp. 97–101.

[155] E. Barth, *Einführung in die Physiologie, Pathologie und Hygiene der menschlichen Stimme*. G. Thieme, 1911.

[156] H. Stern, "Klinik und therapie der krankheiten der stimme," *Mschr Ohrenheilk*, vol. 58, pp. 1–53, 1924.

[157] D. Weiss, "Der begriff des funktionellen mit besonderer berücksichtigung der sprach-und stimmheilkunde," *Mschr Ohrenheilk*, vol. 68, pp. 830–832, 1934.

[158] H. Gundermann, *Die Berufsdysphonie: Nosologie der Stimmstörungen in Sprechberufen unter besonderer Berücksichtigung der sogenannten Lehrerkrankheit*. Thieme, 1970.

Publications

International journals

[J1] Szaszák, Gy., **Tulics, M. G.**, & Tündik, M. Á., "Analyzing FO discontinuity for speech prosody enhancement," *Acta Univ. Sapientiae Elect. Mech. Eng*, vol. 6, no. 1, pp. 59–67, 2014. (6/3=2 points)

[J2] **Tulics, M. G.**, & Vicsi, K., "Automatic classification possibilities of the voices of children with dysphonia," *Infocommunications Journal* Vol. X. No.3. pp. 30-36., 7 p. 2018. (4 points)

[J3] Kovács, A., **Tulics, M. G.**, Tündik, M. Á., Moró, A., Gróf, A., "Magmanet: Ensemble of 1d convolutional deep neural networks for speaker recognition in hungarian," *Phonetician*, vol. 115, pp. 72–86, 2018. (6/5=1.2 points)

[J4] **Tulics, M. G.**, & Vicsi, K. (2019). "The automatic assessment of the severity of dysphonia," International Journal of Speech Technology, 1-10. (6 points)

[J5] Szántó, D., Jenei, A. Z., **Tulics, M. G.**, & Vicsi, K., "Developing a Noise Awareness Rising Web Application within the "Protect your Ears" project," *Infocommunications Journal* 2020. Accepted

Hungarian journals

[J6] Sztahó, D, Kiss, G, **Tulics, M G**, Czap, L, Vicsi, K, "Számítógéppel támogatott prozódiaoktató program," *Alkalmazott Nyelvészeti Közlemények* 9 : 1 pp. 144-153. , 10 p. (2014) (2/4=0.5 points)

International conferences

[C1] **Tulics, M. G.**, Kazinczi, F., & Vicsi, K., "Statistical analysis of acoustical parameters in the voice of children with juvenile dysphonia," In *International Conference on Speech and Computer* (pp. 667-674). Springer, Cham. 2016. (3/2=1.5 points)

[C2] **Tulics, M. G.**, Szaszák, Gy., Mészáros, K. & Vicsi, K., "Artificial Neural Network and SVM based Voice Disorder Classification," In *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE. 2019. (3/3=1 points)

[C3] **Tulics, M. G.**, Lavati, L. J., Mészáros, K. & Vicsi, K., "Possibilities for the automatic classification of functional and organic dysphonia," In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2019. (3/3=1 points)

[C4] **Tulics, M. G.**, & Vicsi, K., "Phonetic-class based correlation analysis for severity of dysphonia," In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 000021-000026). IEEE. 2017. (3 points)

[C5] Kiss, G., **Tulics, M. G.**, Sztahó, D., Esposito, A., & Vicsi, K., "Language independent detection possibilities of depression by speech," In *Recent advances in nonlinear speech processing* (pp. 103-114). Springer, Cham. 2016. (3/4=0.75 points)

[C6] Sztahó, D., **Tulics, M. G.**, Vicsi, K., & Valálik, I., "Automatic estimation of severity of Parkinson's disease based on speech rhythm related features," In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 000011-000016). IEEE. 2017. (3/3=1 points)

[C7] Sztahó, D., Kiss, G., **Tulics, M. G.**, & Vicsi, K., "Automatic Separation of Various Disease Types by Correlation Structure of Time Shifted Speech Features," In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)* (pp. 1-4). IEEE. 2018. (3/3=1 points)

[C8] Sztahó, D., Kiss, G., **Tulics, M. G.**, Dér-Hajduska,B. & Vicsi, K., "Automatic discrimination of several types of speech pathologies," In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2019. (3/4=0.75 points)

[C9] **Tulics, M. G.**, Szaszák, Gy., Mészáros, K. & Vicsi, K., "Using ASR Posterior Probability and Acoustic Features for Voice Disorder Classification," In *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE. 2020. Accepted

Hungarian conferences

[C10] **Tulics, M. G.**, Jászai, H., & Vicsi, K., "A diszfónia súlyosságának automatikus becslése, a szakértői értékelések szubjektív jellegének figyelembevételével," In: *Vincze, Veronika (szerk.) XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)* Szeged, Magyarország : Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 206-218. , 13 p. 2018. (1/2=0.5 points)

Total publication score: 24.2 points

Independent citations: 13