INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Sistema para Avaliação e Feedback de Manobras Respiratórias Forçadas em Crianças

BERNARDO ROCHA PINHO
Julho de 2018

POLITÉCNICO
DO PORTO

System for Evaluating and Providing Feedback of Forced Respiratory Manoeuvres in Children

Bernardo Rocha Pinho

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Sistemas Computacionais

Orientador: Paulo Gandra Sousa

Co-Orientador: João Pedro Teixeira

Supervisor: João A. Fonseca

Porto, Julho 2018

Resumo

A asma é a doença crónica mais comum em crianças, não curável mas tratável, e é uma causa frequente de admissões hospitalares evitáveis. Devido ao custo associado a equipamentos clínicos para uso pessoal e à crescente universalidade de *smartphones*, tem-se vindo a desenvolver trabalho na extração de parâmetros clínicos de manobras respiratórias forçadas com o auxílio destes últimos, particularmente através do microfone. Contudo, não têm sido explorados métodos para avaliação de qualidade destes registos áudio, limitando a utilidade destas soluções sem a presença física de profissionais de saúde.

Como tal, no âmbito desta tese pretendemos desenvolver um sistema para avaliar três critérios de qualidade em espirometria móvel com recurso ao microfone. Adicionalmente, o sistema foi também utilizado para fornecer *feedback* ao utilizador sobre como melhorar execução da manobra respiratória. Planeamos duas fases de colheita de dados num ambiente controlado, a primeira para permitir o desenvolvimento dos algoritmos de avaliação automática dos critérios de qualidade, e a segunda para validação do feedback associado a cada um.

Na primeira fase foram efetuadas 400 gravações de espirometria com o microfone de um *smartphone* em 55 crianças (dos 8 aos 10 anos), usando um jogo móvel desenvolvido para esse propósito. Um profissional de saúde seguidamente procedeu ao registo da qualidade dessas gravações e a dar *feedback* às crianças. Para dois dos três algoritmos desenvolvidos obtiveram-se correlações relevantes com a classificação do profissional de saúde.

Na segunda fase realizaram-se 93 gravações em 20 crianças, de forma idêntica à primeira fase, mas com o próprio jogo a fornecer o *feedback* em vez do profissional de saúde, para efeitos de validação desse *feedback* automático baseado nos algoritmos desenvolvidos. Verificou-se que todas essas crianças foram capazes de perceber como melhorar a sua execução sem intervenção do profissional de saúde, satisfazendo os critérios antes de atingir o número máximo de tentativas recomendados na literatura existente.

Os resultados obtidos reforçam a ideia de que a Espirometria Móvel, utilizando um microfone, é um campo de estudo viável e promissor. Contudo, estudos com uma maior amostra de crianças representativa de toda a demográfica alvo serão necessários para melhorar o sistema, bem como para garantir a aplicabilidade destes resultados na população geral.

Palavras-chave: Asma, espirometria registada em microfone, avaliação de qualidade espirométrica, pediatria

Abstract

Asthma is the most common chronic disease in children, uncurable but treatable, and is an important cause for preventable hospitalizations. Due to the cost of spirometers for personal use and the smartphone ubiquity, work has been developed for extracting clinical parameters from spirometry manoeuvres using smartphones, particularly through the embedded microphone. However, quality evaluation of those audio captures has not been explored, limiting the usefulness of these solutions in the absence of healthcare professionals.

Therefore, we developed a system to evaluate three quality criteria in smartphone spirometry, and to provide feedback on how to improve users' performance. We recorded 400 microphone spirometries in 55 children (8-10 years), using a mobile game engineered for the purpose, and a healthcare professional registered its quality.

For two of the developed algorithms we were able to attain relevant correlations with the healthcare professional's classification. In a second assessment for the validation of the automatic feedback, we observed that all participants understood how to improve without intervention by the healthcare professional, and were able to satisfy the criteria before reaching the maximum attempt count defined in existing literature.

These results suggest that child microphone spirometry in the absence of a healthcare professional is feasible.

Keywords: asthma, microphone spirometry, spirometry quality assessment, paediatrics

Acknowledgements

I am grateful towards my parents, who have provided a stable environment in which I was able to flourish. Grateful to my dear Marta for encouraging and supporting me along the way, as well as my friend Ismael, who were with me through the very thick and the not so thin, and yet chose to persevere with me. To my friend Texas who has been with me for so many years,

Grateful to my company Medida and Prof. João Fonseca in specific, for treating their employee so well, and this played no small part in the maintenance of my mental sanity throughout this ordeal. Also to my co-workers, members and partakers of the noisy office nº6. For bearing with me and for being always willing to lend a helping hand, especially Rute and Cristina, but also Rui, Ana and Rita. A very special thank you to Filipa, for the prompt availability in spending several tiresome mornings doing data collections with me.

Grateful to Colégio das Escravas, particularly Rita, who enabled us to have easy and convenient access to so many children, for their hospitality and eagerness to help.

Grateful to Victor Chelaru, for all the nights he spent chasing after FlatRedBall bugs with me, without which help I would not have had time to accomplish what I have in this thesis. For everyone on gitter chat, especially Dom and Jared, who kept me company during the haunted late night hours on an otherwise empty campus. For Jan, who offered himself to develop the game assets for Destructoid.

The methods developed in this work are part of Project Personal Health Empowerment (16040) from ITEA 3 Call 3 (ITEA 3 is a EUREKA strategic ICT cluster programme), on the scope of the international consortium of which Medida is a member.

Index

List of Figures

List of Tables

Acronyms and Symbols

Acronym List

BEV	Back Extrapolated Volume
FS	Forced Spirometry
FVC	Full Volume Capacity
FEV_1	Full Expiratory Volume in 1 second
PEF	Peak Expiratory Flow
TIFF	Tiffeneau index
EV	Extrapolated Volume
ATS	American Thoracic Society
RMS	Root Mean Square
FFT	Fast Fourier Transform

1 Introduction

1.1 Problem

Asthma is the most common chronic disease in children. Performing pulmonary function testing typically not only requires specialized equipment, but also specialized medical personnel to determine the validity of a required maximal effort exhalation manoeuvre. It must obey certain established quality criteria, or it will compromise the validity of the end results.

Work has been developed to enable accurate estimation of medical parameters in these manoeuvres outside of a clinical setting, making use of smart devices due to their growing popularity, leveraging their embedded sensors.

Specifically, a project called “Ar.Cade” aims to create a solution to enable monitoring of asthma in children on the absence of their respective healthcare professional, serving as an early warning system for impending asthma attacks. Children interact with this system via a virtual pet game, where the player is meant to perform mini-games with the virtual pet with their breath sound being captured by the microphone for analysis purposes.

However, there lies a problem common not only to Ar.Cade, but to all currently known solutions for smartphone spirometry: assuring the validity of the manoeuvre in the absence of a specialized healthcare professional, limiting the usefulness of the results obtained in these circumstances. This specific problem remains unexplored to the best of our knowledge.

1.2 Objectives

Therefore, in this thesis we propose to develop a system that implements two major features:

- Capability of analysing the audio of a forced spirometry, captured by a smartphone’s microphone, to classify a sub-set of manoeuvre quality criteria;

- Capability of providing useful feedback to children in case some criteria is not met, enabling them to understand how to perform the manoeuvre correctly by themselves next time.

Further, given the nature and purpose of Ar.Cade's virtual pet mini-games, we propose to develop one of them serving as the user-interface for this system, aiming for children between the ages of 5 to 10 years old.

However, we will not deal with the accurate estimation of clinical parameters and the classifying of the user's lung function (i.e. how sick is the user), nor will we focus on developing any other advancements to the game aspect of the Ar.Cade project aside from the mentioned mini-game.

1.3 Performance evaluation

The development of algorithms for forced respiratory manoeuvre quality classification required that we first performed data collection events for these manoeuvres, with the quality of each being labelled by a specialized healthcare professional.

After the development stage, the capabilities of the classifier were evaluated and discussed through its agreement with the healthcare professional (Matthew's Correlation Coefficient), with the main question being: "is the system as good as the healthcare professional in assessing these quality parameters?"

For the evaluation of the feedback loop, a second data collection event will be performed on the same children of the first event, this time however with the quality classification algorithms implemented and appropriate audio feedback for each failed criterion. By observing the children's behaviour between each manoeuvre without any human guidance, we will try to answer the question: "can the system remind a previously trained child of how to perform the manoeuvre correctly, without the intervention of a healthcare professional?"

1.4 Contributions

In attempting to achieve these objectives, several contributions were made:

- Design of new quality criteria for microphone specific lung function testing applications;

- Implementation of a processing algorithm to estimate medical curve shape from an audio signal in mobile smart devices;

- Development of an automatic quality classifier for lung function testing in mobile smart devices;

- Development of a mini-game for Ar.Cade's virtual pet game;

- Full paper accepted for a 30 minute oral presentation: "Automatic Quality Assessment of Smart Device Microphone Spirometry", B. Pinho, R. Almeida, C. Jácome, R. Amaral, F. Lopes, T. Jacinto, R. Guedes, M. Pereira, I. Gonçalves, J. A. Fonseca., SPCS 2018: International Conference on Signal Processing and Communication Systems, July 29-30, Porto.

2 Context

In this chapter we will first present what asthma is and its impact on society, then analyse what spirometry is and how mobile health applications are growing in this sector. This will lay the groundwork for the Ar.Cade project: what it is, where this thesis intersects with it, and the value that it aims to bring to society.

2.1 Asthma

The World Health Organization defines Asthma as a chronic disease characterized by attacks of breathlessness and wheezing that persist over time. Although it cannot be cured, with proper management a good quality of life is possible. (World Health Organization, n.d.)

According to the latest report of the Forum of International Respiratory Diseases (FIRS, 2017), asthma afflicts up to 334 million people. In specific, it is the most common chronic disease in children, and it increases the chance for abnormal lung growth and lifelong respiratory compromise. This is more severe for those living in less affluent countries, in part due to under-diagnosis and consequent under-care.

In higher income countries, it is one of the most frequent reasons for preventable hospital admissions, with some studies accounting asthma as “the reason for more than 30% of all paediatric hospitalisations, and nearly 12% of readmissions within 180 days of discharge” (FIRS, 2017). And as more people are concentrated in an urban environment, the burden on healthcare systems is expected to increase even further (Velickovski et al., 2018).

2.2 Spirometry and Forced Expiratory Manoeuvres

"Spirometry is the most widely used non-invasive test of ventilatory function, (…) suitable for the detection, differentiation and diagnosis of various respiratory diseases, and an objective method for following disease progression or improvement and therapeutic response over time." (Pierce, 2005)

The performance of a forced spirometry (FS) test involves three distinct phases: maximal inspiration; a "blast" of exhalation; and continued complete exhalation to the end of test (Figure 1). This is a manoeuvre that requires the training of the patient by a specialized medical professional to ensure filling the established quality and repeatability criteria(Miller et al., 2005).

Assuring that these criteria are fulfilled is of paramount importance, as neglecting them has led to over 25% of false-positives in diagnosing Chronic Obstructive Pulmonary Disease (Moger et al., 2013), and 50% of false-negatives (Walters et al., 2011). Said criteria will be analysed in detail on the next chapter.

Performing FS tests in children is more demanding, and the technician should have specific training to deal with the situation, also making an effort to create a child-friendly atmosphere (Miller et al., 2005)

On higher end spirometers a plot called a spirogram is generated at the end of each manoeuvre, measuring the flow of air. This is typically presented to health professionals as a volume-time and a flow-volume graph (Figure 2), and allows the extraction of parameters such as:

- Forced Vital Capacity (FVC) – the maximal volume of air exhaled with maximally forced effort from a maximal inspiration, expressed in litres;
- Forced Expiratory Volume (FEV_1) – maximal volume of air exhaled in the first second of a forced expiration from a position of full inspiration, expressed in litres;
- Peak Expiratory Flow (PEF) – the maximum expiratory flow achieved from a maximum forced expiration, starting without hesitation, express in litres per second;
- Tiffeneau-Pinelli index (TIFF) – the FEV_1/FVC ratio.
- Back Extrapolated Volume (BEV) – the total volume expelled at the instant where the line of the volume-time's chart highest derivative intersects the time axis. Measures hesitation at the start of the manoeuvre.

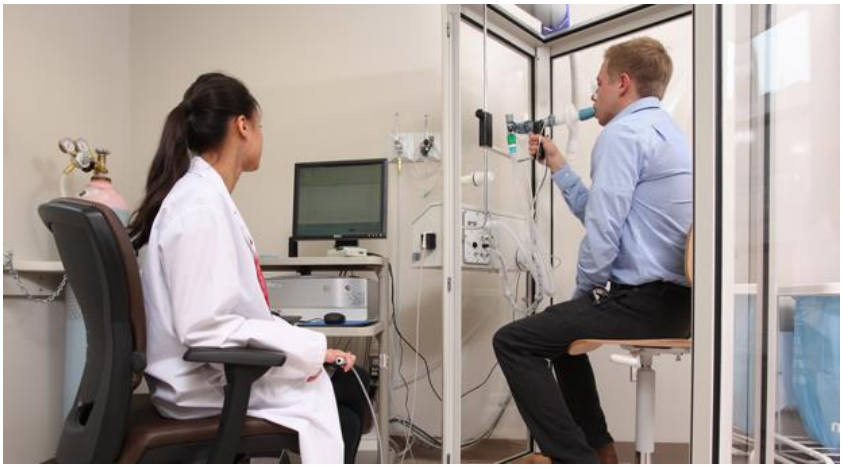

Figure 1- Lung Function Laboratory

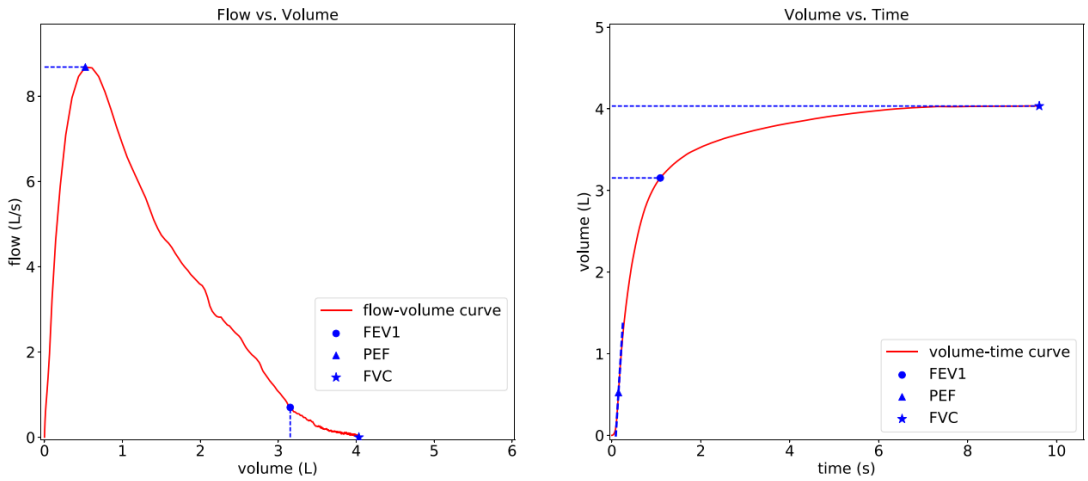

Figure 2 - Flow-volume and Volume-time curves showing the conventional measurements made (Velickovski et al., 2018)

Given the increasing burden of respiratory diseases on healthcare systems and the creation of highly portable spirometry devices, FS testing has begun shifting from exclusive specialists to being available across different healthcare tiers (Derom et al., 2008). Despite being more accessible from a financial standpoint, even the lower end of these portable devices represents a non-negligible investment, with prices ranging hundreds of euros, which leads to the advent of *mHealth* initiatives.

2.3 mHealth

mHealth (mobile Health) is the result of the merging of ubiquitous computing and electronic health systems (eHealth) and is a growing field with recognized potential by the United Nations and the World Health Organization. It is being acknowledged by governments as a "complementary strategy for strengthening health systems". (World Health Organization, 2011)

One of the platforms on which mHealth is growing is the smartphone, which is becoming an increasingly prevalent device in people's everyday lives: Cisco predicts that by 2021, there will be 1.5 smart devices per capita worldwide, having grown from 46% in market share to 54% from 2016 to 2017 compared to non-smart devices ("Cisco Visual Networking Index," 2017)

Despite this fact, currently available asthma related mobile apps mostly serve for logging disease related data, education or for notifications (i.e. medication reminders), and do not attempt to provide automated monitoring capabilities outside of a clinical setting (Tinschert et al., 2017).

2.4 Ar.cade

In an attempt to fill the mobile asthma monitoring applications gap, Prof. João Fonseca, head of the Allergy Units at both CUF Porto Hospital and Institute and founding partner of the Medida company, started the development of the Ar.Cade project. The purpose of this project is to create an ecosystem that enables lung capacity monitoring in young asthmatic children, in the absence of a specialized healthcare professional and of specialized equipment, capable of providing data to the parents and healthcare providers useful to their decision making.

There are 3 components in this ecosystem:

- Mobile virtual pet game: this will be responsible for audio data capture of the child's forced spirometry manoeuvres, providing a game loop that will incentivize the child to perform them well and consistently through time, reporting the data of their asthma status to a server.

- Healthcare professional's back office: this will allow each healthcare professional to access the data of each of his patient's recordings and general historic data in terms of asthma progression over time, allowing to warn the parents to bring the child in for a consultation if an attack seems imminent.

- Parent's companion app: this will allow the parents to monitor aspects like how consistent their child has been in performing the relevant mini-games for monitoring purposes. They will be enabled to intervene on the game world in some way,

encouraging the child to continue. They will also receive reports or alerts from their child's healthcare professional.

Focusing on the monitoring aspect of the virtual pet game, it should be done only using sensors that are universally present in smartphones and tablets, with special emphasis on microphone usage as a game input and data source for medical analysis to reduce adoption barriers. Aside from this, local processing is a requirement of the project to minimize scalability issues.

The game revolves around the concept of taking care of an asthmatic baby dragon that lives on a small planet (Figure 3), and that travels to other planets to perform special activities – mini-games. These mini-games will use the microphone as the main game input, recording the child's FS, analysing it and giving feedback to the child, his/her parents and the assigned healthcare professional on how the child's asthma is progressing.

The mini-game developed in this thesis will be integrated as one of Ar.cade's mini-games. It was developed using the FlatRedBall engine (Chelaru, 2016), a mainly 2D game development engine that runs on C# and that allows for easy cross-platform deployment, including Android and iOS.

Figure 3 – Current Ar.cade virtual pet game prototype

2.5 Value Analysis

Given that the technology and software to be developed in this thesis is meant for integration with Ar.cade, which is attempting to become a commercially viable project, we will analyse the value proposition in this broader scope.

Ar.cade's conception can be framed within the 5 stages of the New Concept Development model (Belliveau et al., 2004). This model was the result of an attempt to define a common language for the conceptualization of innovative projects, up until the point of the actual start of its development. These 5 stages consist of the following:

1. Opportunity identification

This consists of identifying opportunities in the business and/or technological realm worth pursuing in more or less structured methods. These can range from formal systematic tools such as the fishbone diagram, market research ad future scenario planning, to ad hoc approaches like brainstorming and water-cooler conversations (Dewulf, 2013).

In this case, the opportunity was identified mainly through market research, where a technological and medical gap became apparent when looking at the current offerings of off-site monitoring solutions. These either 1) often provided unreliable data to the health care professional, especially in the case of children, and/or 2) presented themselves with a significant barrier of entry due to requiring, or consisting of, specialized equipment that would have to be purchased for that single use case. This being coupled with the fact of the continued growth of smart devices in their pervasiveness worldwide (increasing the average person's access to sensors and substantial computational capabilities) showed an opportunity worth looking into.

2. Opportunity analysis

This second activity involves a deeper analysis of the identified opportunity for confirming that it is worth pursuing. To this end, the same methods used for opportunity identification can be used, although in greater detail.

To further analyse this identified opportunity, two methods were used:

- For the technological side, an initial roadmap draft was made. This consisted of the main technological features required to better understand the research and development costs involved.

- For the business side, market segment analysis was performed, to help further understand the needs of each type of potential customer:

    - Healthcare providers (clinics and hospitals) – these are for-profit businesses interested in maximising patient acquisition, retention and appointment frequency, while minimising associated overhead costs and expenses.

    - Healthcare professionals – these are individuals employed by the healthcare providers, interested in improving their patient's general health and in minimising the costs and effort involved in doing so.

    - Families – these are the children's parents or legal guardians that share the healthcare professional's interests regarding their children.

    - Healthcare payers – these can be a national healthcare systems or insurance companies. These are interested in maximising health systems coverage while minimizing costs related to maintaining their users' health levels.

3. Idea genesis

This is an iterative and evolutionary process where the development and maturation of concrete ideas occurs. Formal tools like brainstorming sessions can be used, but informal processes like unexpected user requests or results from experiments may also shape the ideas generated.

In Ar.cade, brainstorming sessions were used extensively in a multidisciplinary setting, with engineers, healthcare professionals and business entrepreneurs.

4. Idea selection

After generating ideas, it is necessary to choose which to support given the existing time and budget constraints. This is done by evaluating which ideas maximise value to cost ratio typically using formal processes, such as portfolio methodologies and formal idea selection processes with prompt feedback to the submitters.

The latter was used, with the project leader maintaining close contact with each team member and idea submitter, evaluating the technology and software development costs involved for each, and marketability difficulties.

5. Concept and technology development

As the final stage of the new concept development model, here a sufficiently strong case must be made for the business or technology proposition, with qualitative and quantitative information such as objectives, risk factor analysis and size of opportunity.

With the information that was able to be gathered and given the heavy research and development bias of the company, with a propensity toward higher risk projects, the project was accepted.

2.5.1 Value Proposition

This project consists of the already mentioned mobile virtual pet game, which will allow the sending of audio recordings to a back-office available to the child's healthcare professional. This back-office will not only allow access to patient's recordings and corresponding asthma medical parameters, but also to review automatically generated asthma attack alerts and to alert the child's family, should that be desired.

The back-office will be installed at the healthcare provider's level, and the act of linking the device to a specific back-office provides added value in the form of helping to establish customer loyalty.

Aside from this linking, the back-office will also be used to perform a calibration of the device for the child that will use it, in the presence of the healthcare professional. This calibration will be performed in the form of a short tutorial game, which will serve the purpose of introducing

the child to the breath game-mechanic, where the professional will teach the child the specifics required to perform a FS test with the mobile device.

This calibration is required for the obtained data to be relevant. This is not only because the actual microphone specifications of each device can vary from model to model, producing different outputs for the same input, but also that each child may have different standards of what is normal for them. This would affect the outcome of automatic manoeuvre quality classification.

The target market segment will be that of healthcare providers, even though other possible business models were found possible. A flat fee is planned for back-office installation, followed by a monthly-fee based on user count. Each segment's benefits and sacrifices will now be presented.

At the healthcare professional level, the benefits would be that of increased monitoring and crisis early-warning capabilities, enabling improved patient healthcare management. The cost would be learning to use the back-office system, the added timewise burden of calibrating each child patient's device. Overall increased patient count would also be a potential benefit and sacrifice at the same time, as being able to improve the health of more people could produce self-actualization at the cost of having to manage more patient cases at any given time.

At the family level, the benefits would be an increase in ease of mind over the child's asthma condition, with the potentially negative point of an increase in customer loyalty, as their choices would then be limited either to that specific healthcare provider or to others that also use it, should they desire to keep using it.

At the healthcare payer level, if the solution could quantitatively prove itself to be useful in minimizing appointment and asthma crisis incidence rates, this could potentially result in a decrease in insurance premiums for families that were to use it. In turn, this would drive up demand for healthcare providers that would use this solution.

For healthcare providers, the benefits of adopting this solution would be a competitive advantage over other healthcare providers that would not use it, and an increase of customer acquisition, loyalty and satisfaction, assuming a successful solution. Aside from the fees involved, this would come at the cost of potentially lower appointment rates, should the off-site monitoring prove successful. However, this appointment rate reduction could be offset by an increase in the number of patients a single doctor could manage, potentially balancing out the appointment reduction per patient and increasing their overall market share among competitors.

2.5.2 Business Model Canvas

To summarize the business plan, a business model canvas is shown on Figure 4. Sections of the canvas that have so far not been explained will now be detailed.

On customer relationships and channels, it is expected that for this customer segment we would be dealing mostly with groups of healthcare providers, not so much with individual hospitals and clinics as these are becoming less prevalent. This means that the overall number of potential clients is even further reduced, compared to the family market segment for example. This makes each potential client more valuable. In practice, this means that a more personalized relationship is not only possible but advised, such as custom solutions for back-office installation being offered if needed, and 24/7 support available as standard.

Contact channels also reflect this line of thought, with direct contact via e-mail or face-to-face meetings being privileged, along with medical fairs and exhibitions, with the website being a fall-back to let them know how to reach us if they so desire.

Regarding key partners, the *Saúde - CUF* healthcare provider group will be used as a showcase for the solution's capabilities, given the close ties that exist with the Medida company, to aid in further client acquisition. Schools will be important for audio data collection events, building up the forced spirometry audio database. This is required for further R&D of classification algorithms, which will be aided by research centres like INESC TEC and CINTESIS, as they already do at present. Lastly, graphics artists will be required to help produce additional content to the game to prevent burnout from children in the long-term.

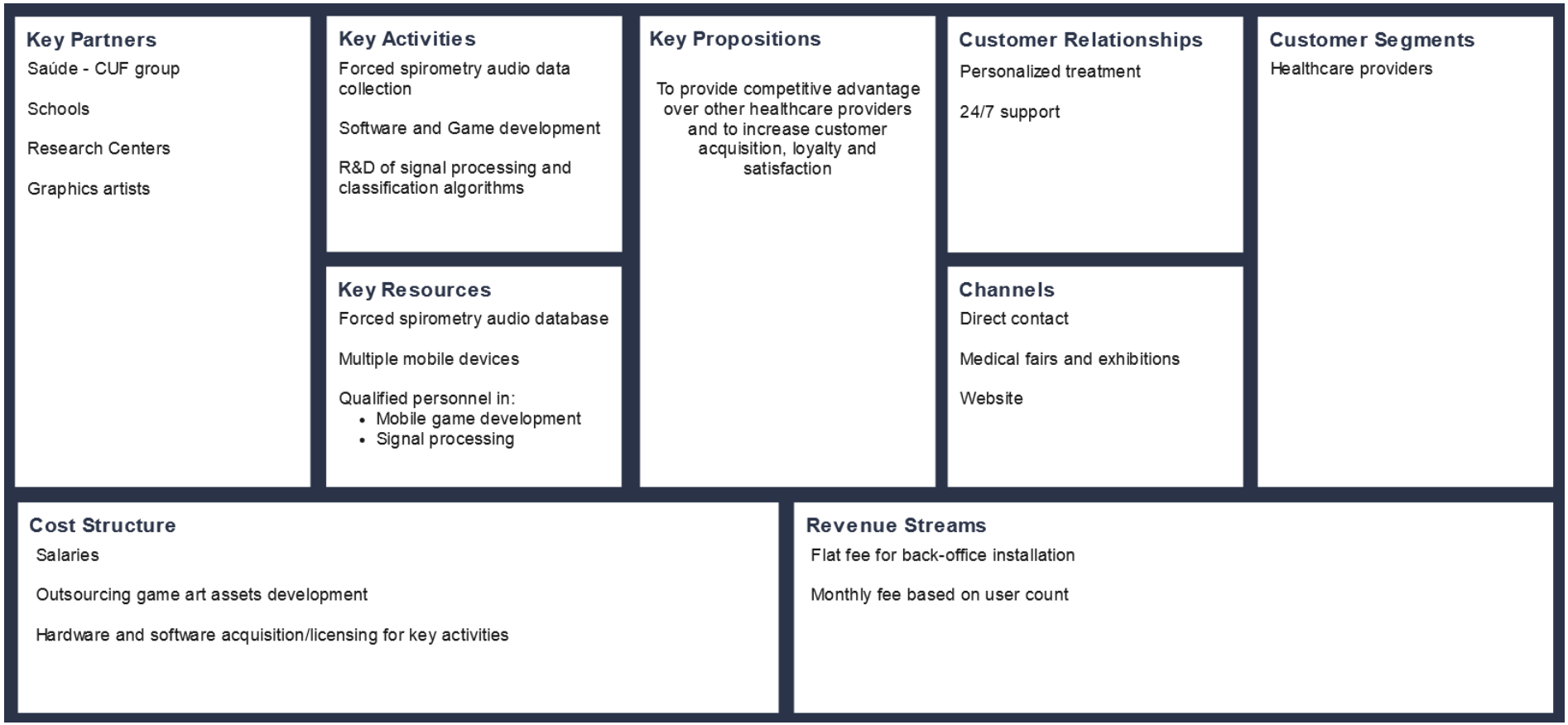

Figure 4 - Business Model Canvas

3 State of the Art

Having shown the context in which this thesis emerges, in this chapter we will provide a literature review concerning two related areas. The first issue involves extrapolation of spirometry curves from audio signals, which are required to validate whether clinical spirometry quality analysis procedures are applicable in our scenario. Afterwards, these quality criteria are presented in detail, along with the state of the art of automatic quality evaluation in clinical and in microphone spirometry.

3.1 Mobile Spirometry Parameter Extraction

Meaningful work in academia has been developed in the field of medical parameter extraction and lung function analysis using mobile smart devices, with or without the use of auxiliary peripherals.

In SpiroSmart (Larson, et al., 2012), the system's goal is to accurately replicate the spirometer's Flow-Volume and Volume-Time curves, in shape and in scale. That is, it attempts to convert the air pressure to actual air flow rate. Data is collected via microphone, and the recorded FS audio data is sent to a server to be processed, with the results being sent back to be presented in the application.

SpiroApp's objective (Stein, 2013) is to provide a simple way to calculate an approximation of a patient's vital parameters, mainly for classification in "good", "maybe bad" and "bad" lung function. It records FS sounds and processes them locally.

mCOPD's system (Liu, 2013) records sound with a microphone and converts air pressure to air flow rate based on said microphone's response model. Although it allows processing of the user's breath sound directly, it is intended that the user perform the manoeuvre with a specially designed sound signal generator, made by a part of a harmonica instrument. Adding the sound generator increases robustness against background noise, since the relevant signal

exists on a narrower range of frequencies that does not coincide with the frequencies of typical background noise. Processing is done locally on the smartphone.

In MobSpiro (Zubaydi, 2016) it is shown a system that attempts to correlate FVC and FEV_1 measurements to the frequency response of human exhalation, as well as estimating the actual flow rate, using local processing and the in-built microphone of a smartphone.

There are smart device solutions that use clinical grade devices as inputs, such as Smart One (MIR, 2014), but those fall outside of the scope of this analysis as they are in practice equivalent to clinical grade mobile spirometers.

3.1.1 Previous Work of the Group

João Teixeira, one of the members of the Ar.cade project, developed a processing pipeline (Figure 5) that allows for the estimation of a flow-time chart from a FS audio capture (Teixeira et al., 2015). The aim of his work was to measure and classify lung function based on signal processing for constructing the flow-time curve, followed by a machine learning stage that enabled the regression of spirometry parameters (i.e. FVC) and the labelling of the illness state. A total of 145 recordings were collected, 101 used for training the models and the remaining 44 samples for testing only.

Each data collection event started with a proper FS test on a spirometer where the ground truth values were noted down for that patient, respecting all manoeuvre quality criteria, followed by recording their FS test using an android app developed for that purpose.

A first stage not represented in the pipeline deals with cropping the audio signal to only include the exhalation sound, detecting the initial instant using the back-extrapolation method that will be discussed further ahead, and the end of manoeuvre with a sliding window algorithm.

On the pre-processing stage, he then proceeds to generate 2 more inputs based off the original cropped signal: the first to compensate for the fact that the patient performed the recording at arm's length from the device to reduce saturation issues, and the second that on top of this converted the airflow measured at the microphone to the one that would be measured at the lips.

Then the envelope generation and post-processing stages transforms each of those inputs into sets of flow-time curves, as shown in Figure 6.

The machine learning stage uses a regression model to estimate clinical measurements. For example, each set of PEF measurements computed from any recording is used as a batch input for the trained regression model, obtaining an estimated PEF value. The process is then repeated for the other types of clinical measurements. The corresponding spirometer measurements acted as ground truth or regression targets.

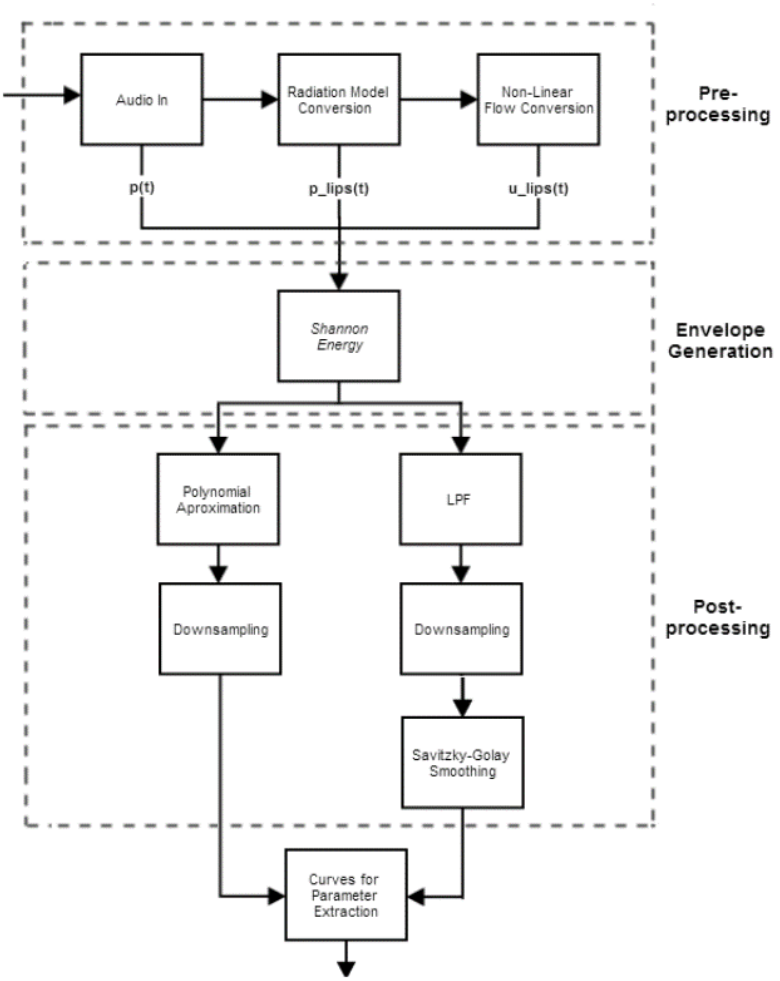

Figure 5 – The last version for the processing pipeline (Teixeira et al., 2015)

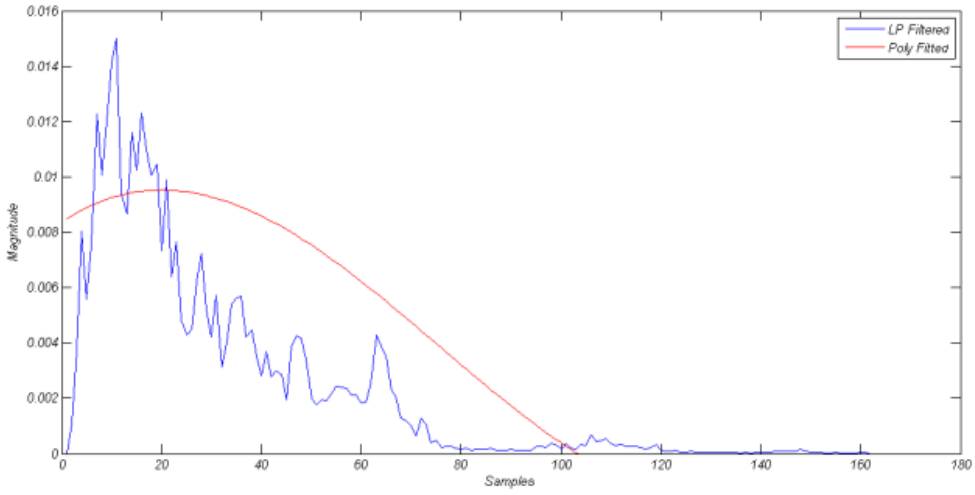

Figure 6 - Example of the resulting two flow-time approximations after processing a FS recording (Teixeira et al., 2015)

As seen on Table 1, results for absolute measurements are substantially worse than the relative measurement that is TIFF, the ratio between FEV_1 and FVC. This would suggest that

there is a scaling issue in the overall estimation of parameters, but that the difference between the spirometer's and the estimate's spirogram in terms of shape is not as significant.

Table 1 - Clinical parameter estimation results of (Teixeira et al., 2015)

Average Error (%)				**Std. Dev. Error (%)**			
PEF	FVC	FEV_1	Tiff	PEF	FVC	FEV_1	Tiff
26.76	32.26	26.95	13.23	25.83	22.66	17.19	12.86

Several limitations have been reported in the work that could affect this:

- Small dataset with little intra-patient samples to properly train the regression model
- Of the samples that were collected, there was a reduced yield of properly executed recordings, in part due to the patient's fatigue in performing a proper FS test on the spirometer beforehand, and the lack of a mouth-piece which added a variable that patients are not used to controlling due to its presence on spirometers.
- The spirometer and smartphone manoeuvres were made separately, so there is no completely reliable ground truth.

It is noteworthy that the future work section in Teixeira et al. acknowledged the need to "devise an algorithm to automatically detect poorly executed FEMs [Forced Exhalation Manoeuvres] to immediately request a repetition during data collection".

Neither his work, nor the works presented before, nor any other works to the best of our knowledge explore the issue of automatic FS manoeuvre quality detection with mobile smart devices, or with exhalation sound in general as the data source.

3.2 Spirometry and Gaming

In young children the difficulty in obtaining a proper FS manoeuvre is higher due to their lower cognitive capabilities, but it has been shown that to incentivize and facilitate the correct FS performance, interactive computer games are used to great effect, providing a beneficial impact especially in young children (Salas et al., 2011).

An example of which can be seen on Figure 7. This game is part of a clinical spirometer like the one shown previously on Figure 2. As the patient exhales through the spirometer's mouthpiece, the bowling ball moves further down the bowling alley in real-time. If the patient succeeds in maintaining their effort throughout the required time, the pins will be knocked down, otherwise the bowling ball will simply revert to its initial position.

Difficulty is adjustable on a simple slider scale (noticeable on the right image), which allows to not frustrate patients to the point of having them give up, especially critical in children, while at the same time trying to push their capabilities to the maximum, as is required.

There are several other games available in spirometers along these lines, such as blowing candles or pushing a ball out of a test tube, but in terms of game design they are entirely identical – they are all physics-based games.

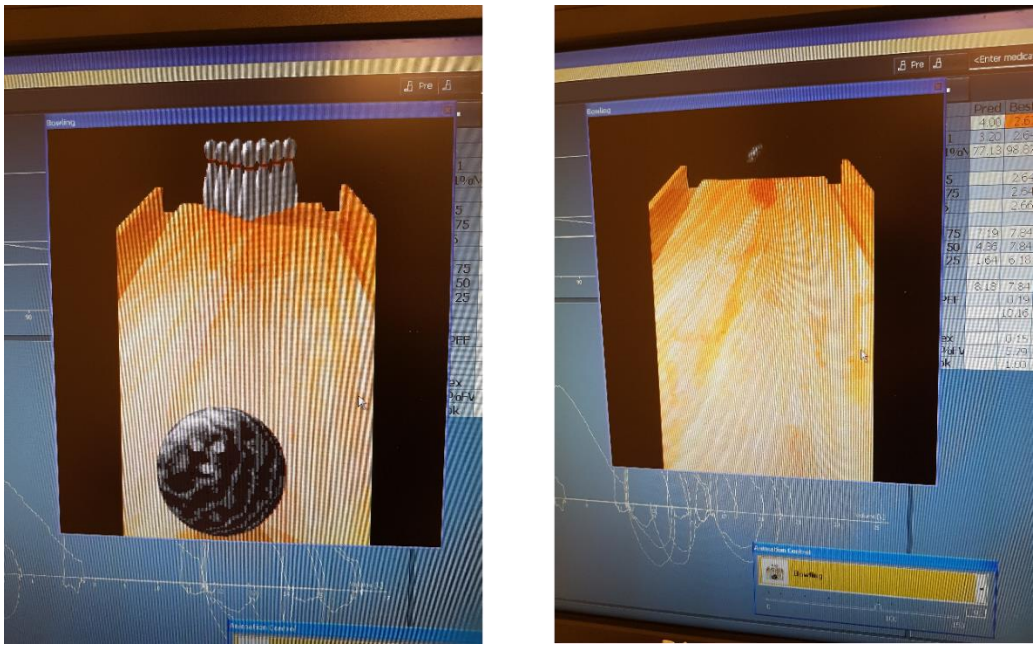

Figure 7 – Bowling ball game used in a clinical spirometer software for FS testing

3.2.1 Previous Work of the Group

André Costa (Costa, 2016) developed a mobile application for this thesis that implemented Teixeira et al. (Teixeira et al., 2015) processing pipeline. The pipeline was ported to Python, creating a standalone audio processing module, and then used in an android app developed in the same language, using an open-source framework called Kivy.

To try and take advantage of this fact, Medida started developing an initial mobile mini-game prototype in parallel using Kivent, a cross-platform game development framework based on Kivy (Figure 8), making use of a physics engine called Cymunk. The author of this thesis was involved in the development of this mobile mini-game. In it, the user would press the red button which would trigger a countdown, after which the microphone would start recording for 4 seconds – enough to make sure the child's breath would be fully captured.

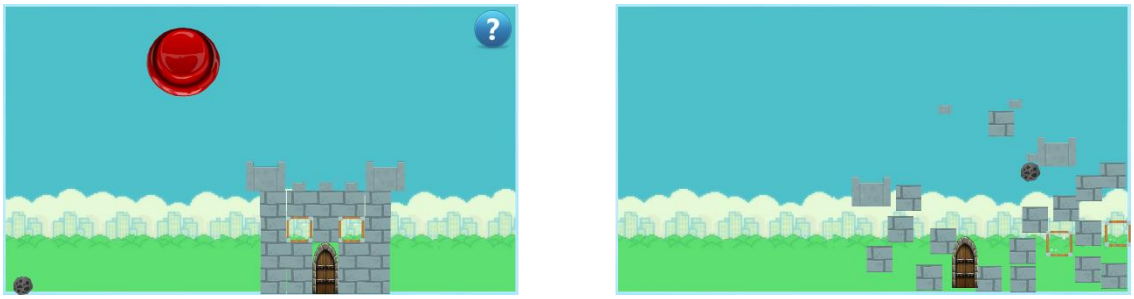

Figure 8 – Kivent mini-game start (left) and destruction (right) phases

During this period, every time the microphone's buffer would be filled with data (approximately every 250ms), the average of the absolute value of its elements would be calculated and used as game input to apply an impulse to the asteroid. Upon impact of the asteroid on the castle, it would in turn destroy it. At the end of the period, the recording would be saved to disk, and the level would reset after all blocks stopped moving.

The game was tested for usability and initial data collection purposes with 16 pre-school children, with ages between 4 and 6 years old (the originally intended demographic of Ar.cade). In a quiet room and each accompanied by one parent, one by one every child was taught how to do a FS manoeuvre, and then proceeded to play the game.

In terms of gameplay, the game was a success as 15 out of 16 children visibly and audibly demonstrated excitement after playing, and 7 out of 9 boys took initiative in asking the name of the game on the app store so they could play at home. In terms of data collection however, those same demonstrations of excitement caused a problem, because they would also happen while the microphone was recording. It was common that children would either interrupt the manoeuvre midway through to:

- Idly watch the castle come apart;
- Express themselves audibly (i.e. laughing);
- Inhale and exhale another time(s), to maximize the destruction of the asteroid upon the castle

Aside from said interruptions, it was also noticeably difficult for the children to understand that they shouldn't be exhaling with pursed lips, but instead with an open mouth, replicating the opening it would have with a spirometer's mouth piece. Even with parental intervention to help them understand how they needed to play the game, it was still not a trivial task to acquire good quality recordings.

Out of 45 recordings, 10 had multiple breaths, 15 were made with pursed lips and 21 had interruptions, totalizing 18 valid recordings. Due to the added challenge of pre-school children in an already non-trivial project, it was then decided to change Ar.cade's focus to an age group with better developed cognitive capabilities as a starting point – 5 to 10 years old.

In terms of game performance, Python proved to be inadequate. Python's *GlobalInterpreterLock* (GIL) prevents multiple threads from executing Python bytecode at once. This means that the capability of multi-threading is severely limited, and even after specific optimizations to reduce the amount of time the GIL is in effect, running the audio processing pipeline while the game tried to show a simple loading animation (a spinning hourglass) made the UI largely non-responsive for 13 to 15 seconds on average.

3.3 Spirometry Quality Analysis

In the quality analysis of FS manoeuvres, there already exist guidelines that are globally well established and recognized. Even though they are designed for proper dedicated spirometry equipment that measure flow-rate directly, it would be unwise to disregard them. Instead, parallels will be established in as much as possible with the reality of spirometry via audio, whose specific challenges will be shown at the end of this section.

3.3.1 ATS and ERS standards

ATS and ERS stand for American Thoracic Society and European Respiratory Society, respectively. In a joint task force they have defined several guidelines to be adopted, including quality control procedures (Miller et al., 2005). In Figure 9, the flow chart describing a high-level view of the process, which will now be explained in detail:

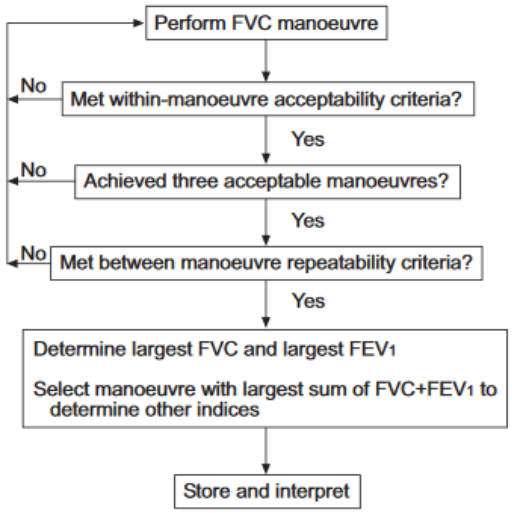

Figure 9– Flow chart outlining acceptability and repeatability criteria are to be applied

An individual spirogram is considered acceptable if it:

- Passes start of test criteria

- Is free from artefacts:
    - Cough
    - Glottis closure
    - Sub-maximal effort throughout the manoeuvre
    - Air leak in the spirometer (i.e. ruptured tube)
    - Obstructed mouthpiece (i.e. tongue on the mouthpiece)
- Passes end of test criteria

3.3.1.1 Start of test criteria

The actual instant at which the FS test is considered to have started is determined by the back-extrapolation method (Figure 10).

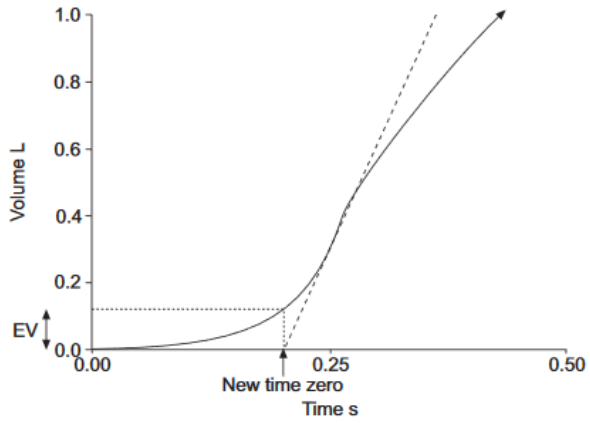

Figure 10 – Back Extrapolation Volume calculation example

This new time zero nt_0 defines the start for all timed measurements, such as the FEV_1. For a computerised approach, the official recommendation is determining the largest slope of the volume-time curve over an 80ms period (Miller et al., 2005). The total volume expelled at the t_{n0} corresponds to the Back Extrapolated Volume (marked as EV on the chart).

According to the ATS/ERS, for a manoeuvre to have an acceptable start of test BEV must be under 5% of FVC, or under 0.150L, whichever is greater. In a polish study in 117 children between the ages of 4 and 10 years old (Tomalak et al., 2008) it was found that this condition was met in 80.4% of them.

3.3.1.2 End of test criteria

Even though the ATS first states that in children the minimum forced expiration time (FET) is to be considered 3 seconds (Miller et al., 2005), they later clarify that "Failure to meet EOT criteria is a prompt to the technician to ensure the subject tries harder on subsequent blows to continue exhaling at the end to achieve the best FVC result(…) their results must not be discarded (…) [and] should be interpreted with the caveat that the EOT criteria were not met" (Miller et al., 2010). And even more recently, the ATS states that it is up to the reviewer to judge whether or not the manoeuvre represents a maximum FVC for that specific patient (Culver et al., 2017).

As corroborating evidence, the polish study (Tomalak et al., 2008) showed that only 23.9% of the children between the ages of 4 and 10 years old were capable of reaching a FET > 3s. Similar results were obtained in other studies as well (Arets et al., 2001) (Müller-Brandes et al., 2014).

3.3.2 Automatic Evaluation Solutions

Currently, and to the best of our knowledge, the only automatic acceptability evaluation solutions that are available to the public only exist in spirometers. Out of all the acceptability criteria that the ATS/ERS has listed, the ones that are commonly used for objective quality control are numerical quality criteria (NQC). In specific, BEV, FET and FVC measurements, with different cut-offs determined for different age groups (Müller-Brandes et al., 2014).

The LUNOKID study involved data from 3133 German children from 4 to 18 years old to determine if these NQC in specific could be used to successfully replace visual inspection of the spirogram, but the study showed that they could not – if only computerised quality control were used, up to 50% of visually unacceptable curves would be accepted.

However, since that study there have been successful attempts at further developing this area of study.

The first that could be found was an algorithm that added 23 new threshold-based criteria to the traditional ones recommended by the ATS/ERS (Melia et al., 2014). The algorithm development was accompanied by and checked against an expert in FS testing, with the new criteria being constructed manually also with the expert's feedback. A decomposition of the flow-volume curve into 5 zones was performed (Figure 11), and specific criteria were applied to each of the zones. As a final output, the algorithm classified a spirogram either as bad (Grade 0), acceptable (Grade 1), or requiring professional judgement (Grade 2).

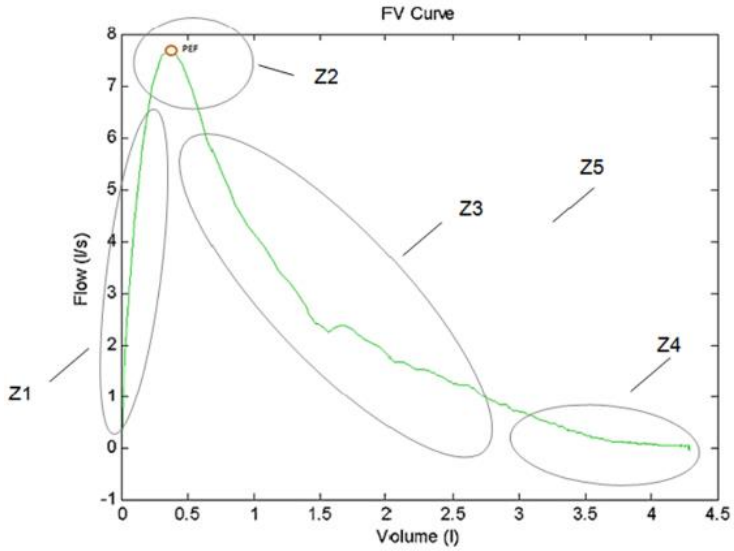

Figure 11 – Spirogram Zones (Melia et al., 2014)

This algorithm produced a classification of 778 spirograms with 96.1% sensitivity (true positive rate) and 94.9% specificity (true negative rate), compared to 67.7% / 75.0% sensitivity and specificity of the equivalent classification using only the traditional criteria. However, there was only one source of ground truth, and the thresholds were manually tuned to increase agreement between the algorithm and the expert on Grades 0 and 1, and to minimize the occurrence of Grade 2 classifications.

More recently, a project using machine learning was developed (Luo et al., 2017). It learned a classification model for early termination, cough, variable flow and extra breath error detection. Features were manually constructed considering previous literature and reviews from experts, totalling 68 per each error. For this project, 19880 spirogram curves were used, being annotated by multiple experts. The classification results are shown on Table 2.

Table 2 - Classification results for Luo's et al. work (Luo et al., 2017)

Error	**Precision**	**Recall**	**F_1-Score**
Early Termination (n=486)	91.5%	92.6%	0.920
Cough (n=446)	81.7%	91.9%	0.865
Variable Flow (n=528)	79.3%	92.8%	0.855
Extra Breath (n=538)	82.4%	87.0%	0.846

3.3.3 Microphone Spirometry

When it comes to microphone spirometry, there exist issues that are specific to it. These are not analysed in any of the studies presented above, nor in any other published work to the best of our knowledge, including the ones on microphone spirometry presented in Chapter 3.1.

On the Kivent game, a common issue observed was exhaling with pursed lips. Although a similar parameter is already analysed in regular spirometry with the mouthpiece obstruction criteria, the end results are opposite of each other: in a spirometer, having a mouth piece obstruction leads to a decrease in measured flow rate. On the other hand, blowing with pursed lips on a microphone leads to higher registered inputs, sometimes even leading to signal saturation issues as were observed on the recordings obtained in the game.

In a spirometer, the tube should be a closed air system with no leaks, with the patient's lips sealing the mouthpiece when performing a FS manoeuvre. In smartphone spirometry, there is no closed system. In (Teixeira et al., 2015) an inverse radiation model is applied to compensate for the distance of the user to the microphone, and an estimate of the air pressure measured at the user's lips is produced. It assumes the user is performing the manoeuvre holding the smartphone at a constant arm's length, using their height to estimate that distance.

With the Kivent game, it was played on a 7-inch tablet that was resting and inclined, with the children being instructed not to touch it while playing, to prevent it from being moved and maintain a more constant distance to the player. However, we observed that the children's common behaviour was to lean back while inhaling, and lean forward not only to exhale but also to more closely observe how the castle fell apart as they exhaled, to the point where their faces almost touched the screen. Trying to instruct them to remain in a more static stance proved to be unfruitful as it is unnatural for them, and even when they tried to remain more static, they would instinctively get much closer to the device as they exhaled. This introduces a non-linear relationship between the true flow rate and the estimated one, whose severity depends on the smartphone sensitivity and the range of motion during the manoeuvre.

A third issue is that of background noise. Spirometers only capture the patient's airflow, but smartphone microphones capture everything around them. This is a specific issue in performing a microphone forced spirometry and, unless the patient is using headphones, there can be no audible incentives. Nonetheless, it is recommended in the official ATS guidelines (Miller et al., 2005) and part of common practice to have the technician verbally and continuously incentivize the patient not to stop exhaling as he nears the end of the manoeuvre.

Lastly, the differences between device's microphones must also be acknowledged. The Kivent game was deployed to 3 android devices, two smartphones and one tablet, for debugging purposes. It was noted that for the same distance to device and manoeuvre effort, the asteroid's speed was very noticeably different which was further verified when analysing the raw audio files. This allowed us to conclude that devices' microphone sensitivity is also a

variable in the equation, and it cannot be easily controlled as that level of granularity in device specifications is not publicly available.

This shows that microphone spirometry presents a unique set of challenges, and while official guidelines should be kept as a starting point, it is important to recognize that the list of possible errors increases with the microphone and that additional quality checks should be analysed.

4 Evaluation on Different Approaches

Given the topics presented, it is useful to discuss possible approaches for three main categories: spirogram curve extraction, the game design concept, and the implementation of the automatic quality analysis module.

4.1 Spirogram Curve Extraction

To achieve the goals of this thesis, the curve extraction step is a stepping stone, not the main purpose. Therefore, given that of all the presented solutions we only have access to the code of Teixeira et al. (Teixeira et al., 2015), and that it can successfully extract a spirogram curve from audio input, from a time efficiency standpoint the initial work will concern the adaptation of Teixeira's et al. processing pipeline.

However, there is an important aspect that must be considered. In their work the parameter extraction step is based on a regression model that requires training. From the research presented in previous chapters, it is known that spirometry in children is not completely equivalent to spirometry in adults, and we are lacking children's FS audio samples of acceptable manoeuvres. This means that by implementing Teixeira et al.'s pipeline (Teixeira et al., 2015), in the current circumstances we will not be able to properly train the regression model, being forced to work in dimensionless or meaningless values where only the shape of the curve has valuable information.

As it was also presented previously, some of the traditional ATS/ERS quality parameters require access to absolute values in Litres or Litres/second. So, by going this route as is, we will only be able to assess quality criteria measured in relative values.

On the other hand, both in mCOPD (Liu, 2013) and in MobSpiro (Zubaydi, 2016) successful attempts were made at ground truth establishment of sound pressure to airflow via a different route, using bladeless fans, an anemometer to measure wind speed, and the smart device's microphone.

Should we be able to replicate their success, that would enable us to resolve this issue. However, scalability issues would arise, not in the scope of this thesis, but in Ar.cade's scope. Should this route be taken, it would have the implication that for the game to be successfully deployed, the user's device would have to be calibrated in the same way, forcing him or the medical expert to acquire the equipment to do so.

While not an insurmountable challenge as the equipment would cost somewhere around 100 to 200€ at the current time of writing, it goes against the ethos and vision of the Ar.cade project. Due to this fact, we will focus on parameters that do not require absolute airflow values and will analyse the results that are possible with them.

4.2 Game Design Approach

Regarding this topic, games that revolve around using the microphone as input are rare, and when it comes to using forced spirometry in specific, to the best of our knowledge we are left with none, aside from the ones that are part of the software suite of some spirometers that we have mentioned on Chapter 3.2.

However, as previously presented, research has shown that those spirometer games are beneficial in incentivizing FS tests. Given the children's feedback received on the Kivent game, it would be reasonable to say that they like these physics-based games with their breath as the input. And again, remembering that the role of the game in this thesis is to provide an incentive for the children, and not an end in itself, it should become clear that as long as there aren't any outstanding usability issues and that the target demographic wishes to play the game, then that shows the game is fulfilling its purpose.

Therefore, we conclude that not only is there a lack of existing alternative solutions, but current knowledge suggests that the Kivent game's conceptual design is satisfactory. This is not to ignore the issues that the Kivent game had; solutions for those will be discussed in detail on chapter 5.1.3.

4.3 Spirogram Quality Analysis

Given that the standard NQC criteria have been shown to be ineffective on their own for spirometry quality analysis, a more robust solution is required. Looking at the different approaches taken by (Melia et al., 2014) and (Luo et al., 2017), it is important to note the advantages and disadvantages on each of them.

As already stated, the algorithm and quality criteria established by Melia et al. were manually built and tuned with the feedback of a medical expert, not through machine learning techniques, and were able to produce good results. Each of these criteria, their thresholds, their weight in the final classification and the algorithm flowchart is described in detail (Figure 12), allowing us to replicate its implementation.

Out of 28 criteria used in the system, 9 are based on absolute value thresholds. These are not evenly distributed, as 7 of these absolute value criteria are concentrated on the classification of Zone 4, that corresponds to validating end of test quality.

Despite this limitation, it makes for an appealing first approach due to ease of implementation, compared to implementing a machine learning method like that of (Luo et al., 2017). In their work, they had a database of 19880 spirograms, manually labelled by several experts.

Although a database of this size and quality may not be strictly necessary to produce any sort of meaningful results, the effort required for this approach was deemed higher with no correspondent increase of positive result expectations, compared to the other approach. This is due to the timeframe available to this thesis for data collection, coupled with the absence of an already existing database of children's FS sounds.

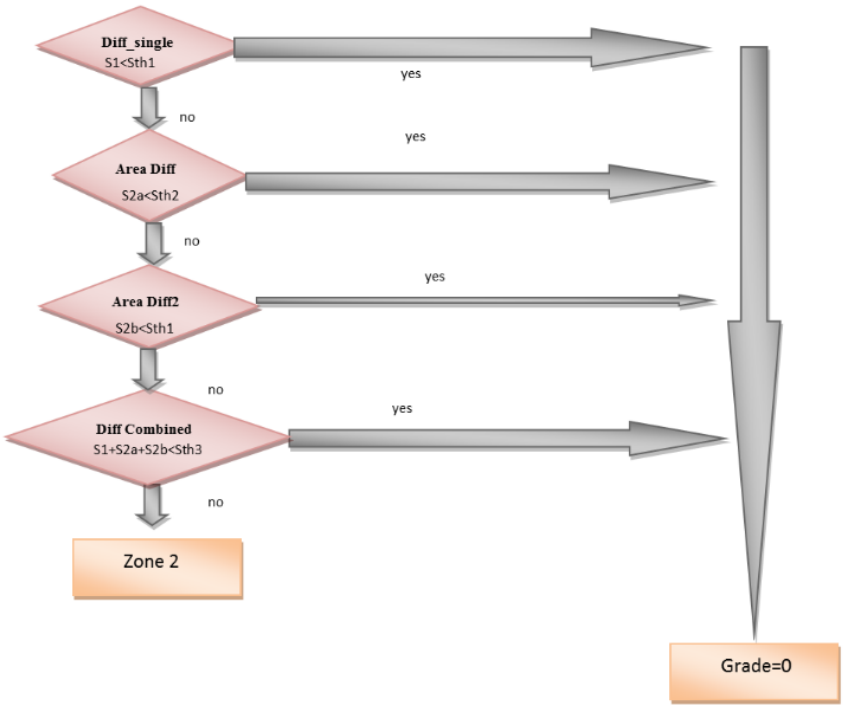

Figure 12 – Flowchart of the algorithm for Zone 1 (Melia et al., 2014)

However, implementing any of these solutions by themselves only allows to begin to try to answer the question of manoeuvre validity as it pertains to regular spirometry. It does not consider the peculiar needs of microphone spirometry, because they were developed for spirometer usage.

This means that an attempt to produce additional specific criteria should be made, either through manual extraction or machine learning or any other method. In fact, in Chapter 6 we will show the steps involved on the manual extraction of features for the development of two microphone spirometry specific criterion.

5 Mini-Game Design and Implementation

Given the children's feedback obtained from the Kivent game, as well as the type of games available in some clinical spirometers, the mini-game consists of an Android physics-based game. It was implemented in C# using the FlatRedBall game engine (Chelaru, 2016), allowing integration with the Ar.Cade project. This game revolves around the use of the smart device's microphone as the main game input. It is able to record the FS manoeuvre and assess if it passes certain quality criteria: in case of a sub-par manoeuvre, the game should be able to instruct the user on how to improve; in case of success, a reward should be given. And finally, it allows for medical input logging.

It is split into three main modules: the Game module (GUI), the Recorder module and the Quality Analyser module. In Figure 13 a high level sequence diagram of the game is presented, showing the main interactions between the different modules, which will now be presented in greater detail.

5.1.1 Game module

This module represents the GUI with which the user will interact with, handling the main game loop. Given the success of the Kivent game in terms of gameplay, we will try to take that formula as a foundation and build upon it, offering solutions to the problems it showed. The Farseer physics engine (Andringa, 2014) was used to enable simulation of collisions without having to develop a custom solution.

5.1.1.1 Visual Feedback Timing

As has been shown previously, physics-based games are a common denominator in clinical spirometer games and were also used on the Kivent game. The visual feedback of the children's

effort in these cases was instantaneous: as they exhaled they could see an immediate and proportional reaction happening in the game.

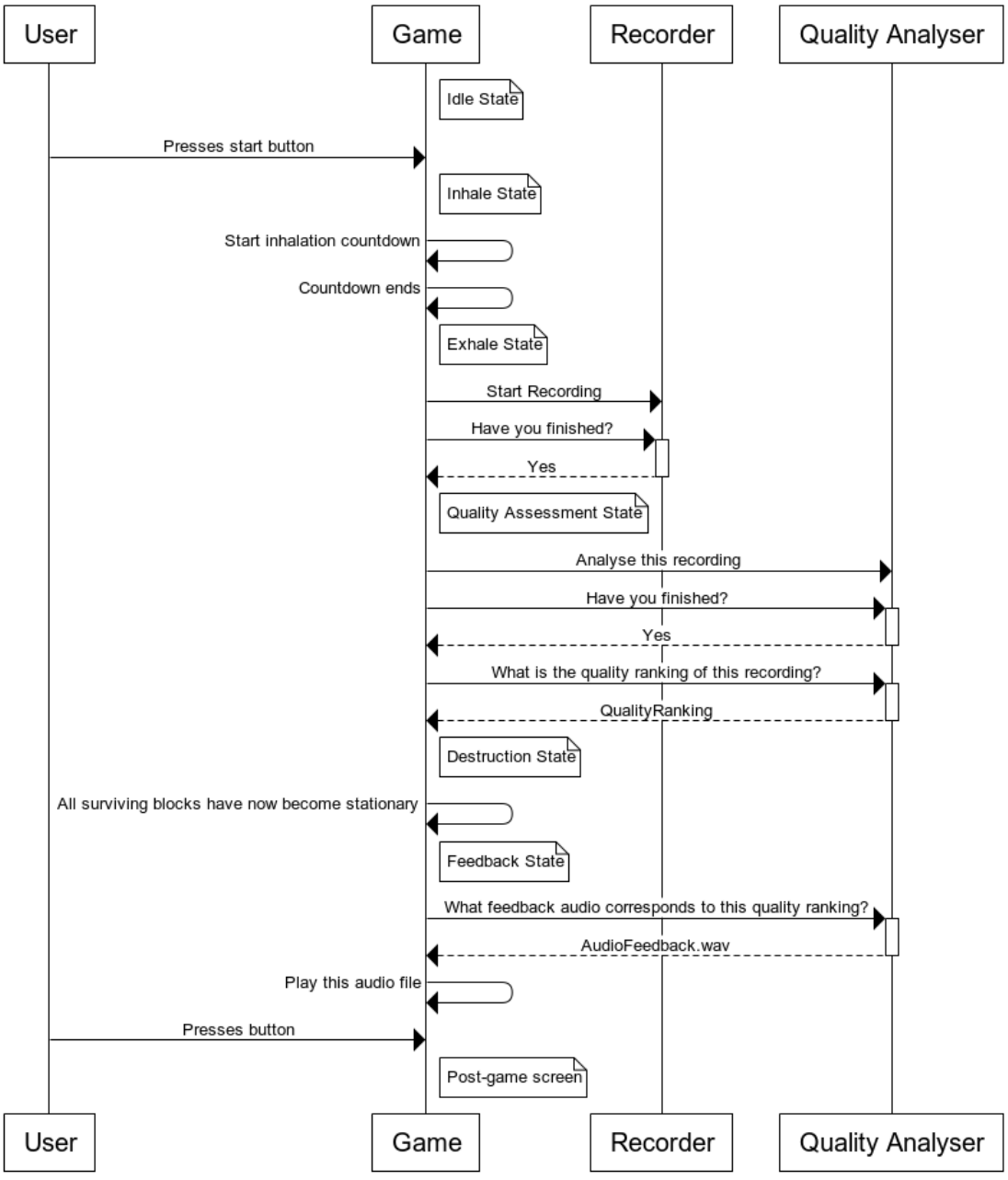

Figure 13 – High-level sequence diagram of the mini-game

However, it is not feasible to validate the manoeuvre quality in real-time. This forces us to make an important choice: should we allow the user to have a meaningful impact on the game before we are able to determine the quality of their manoeuvre?

As was previously stated, it was noted on the Kivent game that children would repeatedly exhale multiple times to maximise the destruction of the target castle, even after being instructed otherwise. Another observed scenario was the involuntary interruption of the

manoeuvre due to the vocalization of their enjoyment with the destruction of the castle. Aside from internal interference such as these, there also exists the possibility for background noise external to the player to have an impact on gameplay.

On that game an unbuffered approach was used, in that the user had an impact on the game that did not depend on the quality of his manoeuvre, but only on the magnitude of the signal measured by the microphone.

On the other end of the spectrum in terms of visual feedback is the buffered feedback approach. In practice, this would mean adding an intermediate stage to the game - i.e. filling up a meter-, followed by a final stage where the result - i.e. the initial velocity of the Kivy game's asteroid - would be dependent on the result of the previous stage.

The advantage of this approach in comparison to the unbuffered one is that it allows to detect any issues with the recorded data before deciding what result should be presented, rewarding the user only in situations where an appropriate input was given.

Although the buffered approach does introduce a new element to be planned and integrated in the game that is simply not required in the unbuffered approach (the intermediate stage), on the latter there exists the specific challenge of coming up with a game design that remains consistent throughout the whole range of potential inputs that may be given. This shows that there are specific implementation challenges to be considered regardless of the chosen approach.

Considering that the mini-game is simply a means to an end, the buffered approach was favoured over a potentially more immersive gameplay experience due to higher gameplay and feedback consistency.

5.1.1.2 Game Loop

Taking cues from both the Kivent game and the Ar.cade concept, we have decided that this mini-game should consist of the baby dragon (called Fred) exhaling a fireball, inciting the player to exhale as he does. This projectile would get thrown at a structure composed of blocks with different materials. The dragon's fireball would have different upgrade states depending on the quality of the manoeuvre, and those upgrades would define which blocks could be destroyed and what sort of animation the fireball's explosion would have, providing different levels of instant gratification to the player proportional to the manoeuvre's quality. This mini-game was given the title of Destructoid - game assets were custom made for it by Jan Halfar, a Czech game designer with over 20 years of experience in the field, as he developed interest in our project. Some of these assets can be seen on Figure 14.

Figure 14 – Destructoid game assets

A simple state machine was used for the main game loop, consisting of five states:

- Idle state (Figure 15)

Here the whole level is in view, Fred and the structure. Aside from this, a button to start the level is also present.

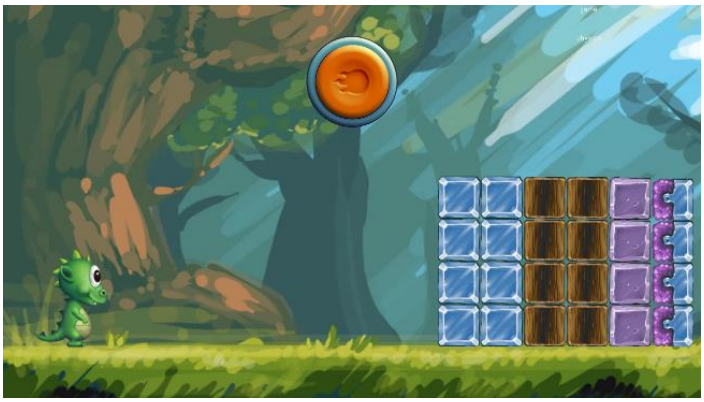

Figure 15 - The mini-game's idle state

- Inhale state (Figure 16)

When the button is pressed, the screen zooms in on Fred to focus the player's attention on the visual cues provided during inhale and exhale (i.e. the dragon's effort and mouth aperture), a countdown with visible and audible feedback starts, giving clear cues to the player on when exactly he should exhale. As the countdown approaches the end, any other background sound effects are gradually muted, so the game itself does not produce interferences on the recording in case earphones are not in use.

Figure 16 - The mini-game's inhale phase

- Exhale state (Figure 17)

After the countdown finishes, audio recording starts and the dragon performs an exhale animation that continues until the end of the audio recording. During this animation, fire particles are blown and a slowly expanding proto-fireball starts to form. This whole process is meant to serve as an incentive for the child to imitate Fred, only exhaling once and for as long as they can. This is done to try and reduce the chance of children attempting to do multiple breaths, but without having to rely on the development of specific quality criteria and feedback for this issue.

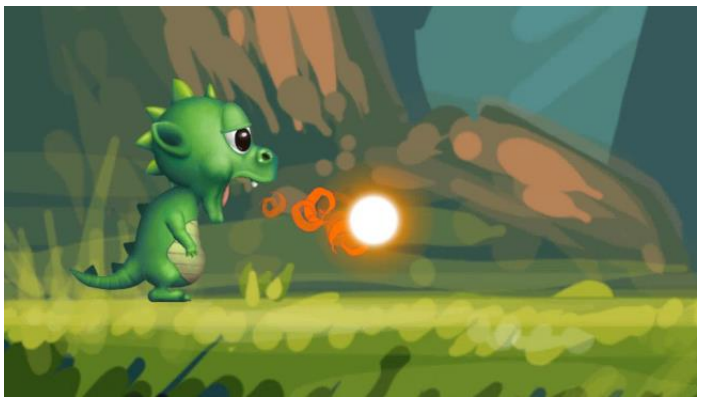

Figure 17 - The mini-game's exhale phase

- Quality Assessment state (Figure 18)

In this stage the recorded audio is sent to the quality analysis module. While the result is awaited, game sounds fade back in and Fred returns to his idle position. Quality should be graded on a scale of 0 to 3 given the number of quality criteria we intend to evaluate, with the projectile's quality of visuals, animations and destruction capabilities being proportional to this rating. again to provide appropriate instant gratification to the player according to their performance:

- 0 stars – If the child exhales with sub maximal effort, the proto-fireball downgrades by cooling down into a simple “air ball”, not causing any destruction as it barely reaches the target;

- 1 star - If the child exhales with higher than minimum effort but with pursed lips, their exhaled air will be cooler. Therefore, the in-game consequence will almost the same as for 0-star ranking: the proto-fireball will cool down into a simple “air ball”, but it will go with enough force to break a few glass blocks;

- 2 stars – If the child exhales with good mouth aperture, their exhaled air will be hotter. Therefore, the proto-fireball will transform into an actual fireball, capable of breaking through several glass blocks but finding higher resistance with wood, and not being able to do anything to stone;

- 3 stars – If the child has good mouth aperture and has an “explosive” exhalation start (good start of manoeuvre criteria), they will receive the fireball’s explosive upgrade and be able to destroy up to stone grade blocks.

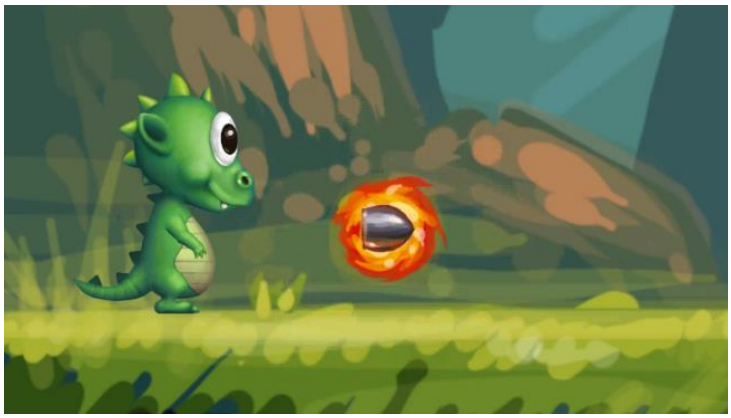

Figure 18 - The mini-game's quality assessment state

- Destruction state (Figure 19)

When the manoeuvre quality is assessed, the projectile receives the appropriate upgrade and is launched towards the structure. The camera tracks the projectile and ends up centred on the structure.

Figure 19 - The mini-game's destruction state

- Feedback state (Figure 20)

After every remaining block returns to a motionless state, an end of level scoreboard appears where the child receives a bronze, silver and golden star depending on the result of their manoeuvre quality assessment. More importantly, they are provided audible feedback on their performance, and in case of a sub-par manoeuvre the feedback will explain how to achieve the next star and break the next type of block. By pressing the button, the game moves to the post-game screen, detailed on the following section.

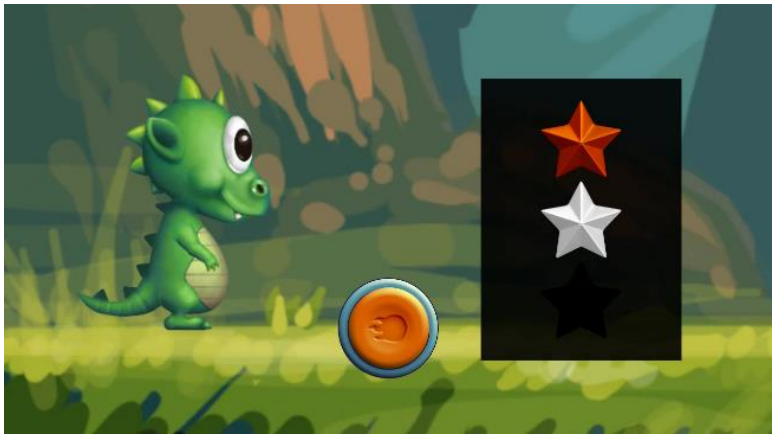

Figure 20 - The mini-game's feedback state

5.1.1.3 Medical UI and Logging

To enable data logging of the children's info we collected, a UI was developed. Specifically, a pre-game (Figure 21) and a post-game screen (Figure 22). These were developed with a healthcare professional user in mind, since they will be the ones interacting with this application for data collection purposes.

The naming convention for the audio files is "patient_<*pid*>_take_<*tid*>_<*guid*>.wav", where *pid* is the player ID as defined on the pre-game screen, *tid* is the ID of this player's recording.

The *tid* is a sequential integer starting from 0. A JSON file is used to store each pair of [*pid*,*ltid*], where *ltid* is the last *tid* used by a specific *pid*, enabling the avoidance of file name conflict.

Aside from the JSON configuration file, a comma-separated value file is used for the storage of information inputted by the healthcare professional, such as manoeuvre classification, and metadata associated with the audio file such as sample rate.

On the pre-game screen, the "Gerar ID" button allows the user to generate a new patient ID that does not clash with any other ID already in use on the current device. It does this by looking at the last used ID which is stored on the JSON file, incrementing that ID and updating the configuration file to reflect that change. In case the healthcare professional wants to collect data from a player that has played on an earlier date and associating him with his previously generated same ID, the "ID" button can be used which opens an input text box. This allows the user to input the previously used ID for that player.

The buttons "sopro frio", "sopro quente" and "sopro forte" reflect the requirements for the first data collection event, as will be described in detail on chapter 6.1. For a version of the mini-game without automated quality assessment, the healthcare professional is guiding the player to perform different kinds of exhalations. These buttons allow the healthcare professional to choose what kind of outcome the player should receive before the level starts, based on the intended type of exhalation. After pressing one of these buttons, it goes into the game screen with the machine state set to idle state. On the version of the game with automated quality assessment, while these buttons are still present they all provide the same outcome, as the outcome is then decided on the quality assessment state.

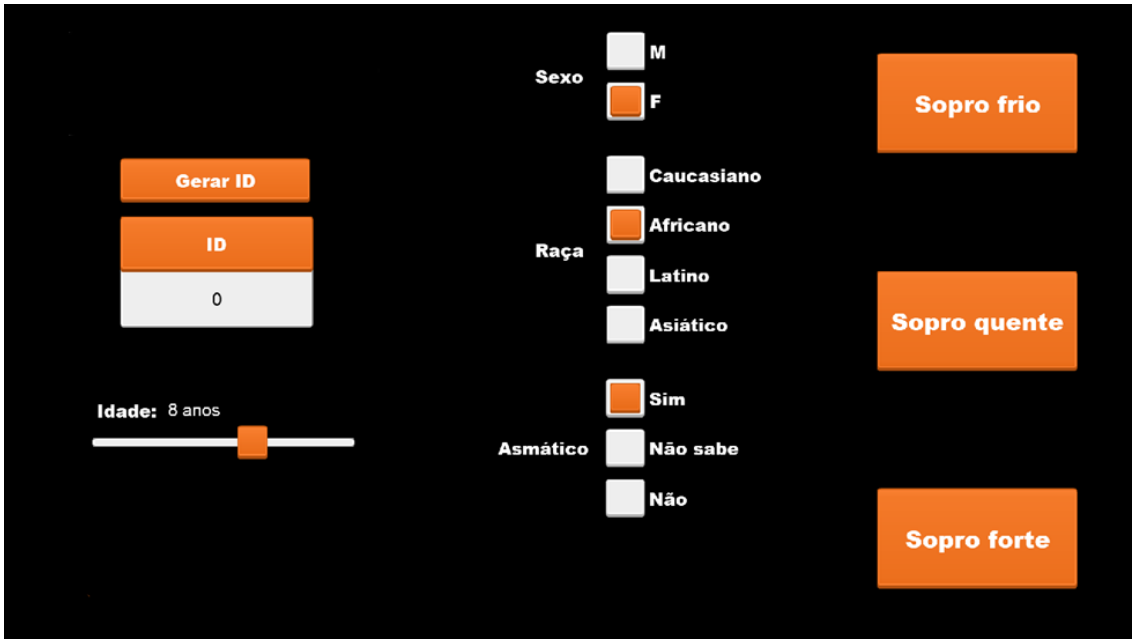

Figure 21 - Pre-game screen

On the post-game screen, we enable the healthcare professional to register the quality of the manoeuvre that just occurred. Though not all categories were strictly necessary for the scope of this thesis, they were nevertheless desired for future work.

The button “Repetir Paciente” allows the user to do another take with the same player without having to re-enter his info on the pre-game screen. The button “Novo Paciente” cleans the info on the pre-game screen for the case of switching to a new player. The button “Guardar e Sair” is there just to provide a means for a graceful exit of the game. Each of these buttons has one feature in common, which is that of creating a new entry on the comma-separated value file with the information that was inputted by the healthcare professional.

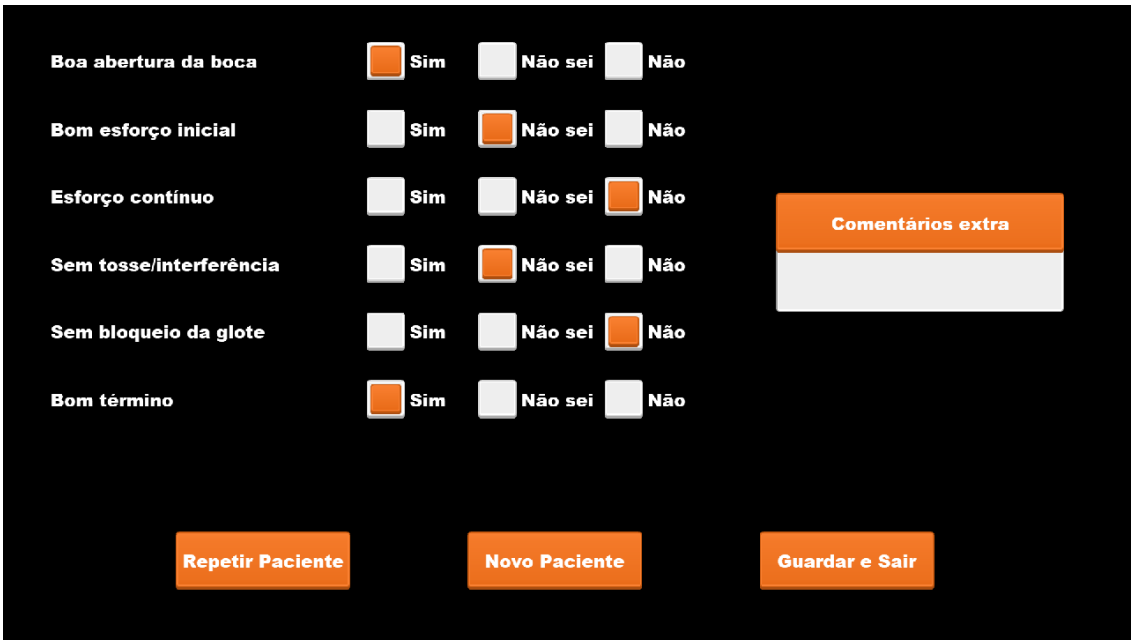

Figure 22 - Post-game screen

5.1.2 Recorder Module

This module is responsible for the capture of audio samples obtained from the device’s microphone, with the purpose of saving each recording to disk. It has low-level access to the raw microphone buffer values, as well as to the recording properties such as sample rate and bit-depth for logging purposes. Given microphone recording is an intensive I/O task, it is performed in a separate thread to avoid performance issues.

As has been discussed in previous chapters, the expected length of proper spirometry manoeuvres in children of this age is less than 3 seconds. To provide for some margin of error, 3.5 seconds were chosen as the initial target for recording length, the same used on the Kivent game. We only had a case where this was not enough, and it was due to the child being extremely shy, and therefore nervous, failing to react adequately to the start of test countdown like all the other children. For edge cases such as this, no reasonable hard-limit for manoeuvre timeout can be determined.

An automatic end of manoeuvre detection was considered, as it would allow for solving edge cases such as this, as well as improved gameplay flow, but was ultimately decided against. The gameplay advantage it would bring is the reduction of the time to move from the game's exhale stage to the processing stage, which in turn reduces the opportunities for multiple breaths to be taken. Comparing to a 3.5 second hard limit, best case scenario would be a saving of around 2.5 seconds.

On the other hand, worst case scenario would be some sort of edge case occurring during the manoeuvre, causing a premature and false detection of its ending. This would force an unexpected invalid manoeuvre classification that would confuse the user, thereby reducing the quality of the feedback loop. Inserting another moving part into an already complex solution would be counterproductive, especially in such an early and experimental stage. The trade-off is then made in favour of a safer and more predictable implementation – a static end of recording length of 3.5 seconds.

The WAVE file format was used for the saved audio file, due to it being lossless: discarding information at this stage would possibly hinder the classification performance of the quality analysis module. The game module provides the intended filename of the generated WAVE file to the recorder.

5.1.3 Quality Analyzer

This module encompasses all the required processing to classify a microphone FS manoeuvre, with the purpose of indicating the appropriate feedback audio file based on the calculated manoeuvre quality. While the development process of the quality analysis algorithms is presented in full on chapter 6, here it is relevant to add that this module also encompasses the implementation of the processing pipeline by Teixeira et al., as one of the criterion depends on the usage of the volume-time curve.

To allow for this, we used the AltaxoCore ("Altaxo," 2015) and MathNet.Numerics ("Math.NET Numerics," 2002) C# libraries to implement the required Savitzky-Golay filter and cubic spline interpolation, respectively.

6 Quality Analyser Development

In this chapter we will present the steps taken to develop the algorithms for each selected quality criteria, starting from the organization of the data collection events, leading to the methodology for the construction of each algorithm and the interpretation of their performance, finalizing with their integration on the Quality Analysis module.

6.1 Data Collection

To enable the algorithm development, an initial data collection phase is required given no access to an already existing FEM audio database in children of our target age group. To gain access to our target demographic healthy children and some children with asthma, we chose to contact a primary school. Specifically, we reached out to "Externato das Escravas do Sagrado Coração de Jesus do Porto".

Written permission requests were sent to the legal guardians of children from the 1st to 4th grades (Annex A), but due to the order in which we received the permissions from the parents we were only able to perform data collection on children from the 3rd and 4th grades – 8 to 10 years old.

The protocol followed for this data collection is presented in detail in Annex B. In summary, we collected the self-reported data of sex, ethnicity, age and if they had asthma. Each child received an internal random ID for same-child recording analysis and comparison. The data collection event of each child was conducted by a specialized healthcare professional, and the following types of manoeuvres were:

- "Cold air" test: Have the child exhale with pursed lips, focusing on exhaling cold air with no requirement of maximal force.

- “Hot air” test: Have the child exhale with a wide open mouth, focusing on achieving a good aperture with no requirement of maximal force.

- Maximal force test: Have the child exhale with a wide open mouth with the added requirement of maximal force, similar to a regular spirometry test.

The first two tests were made to enable the research of ways to distinguish exhaling with pursed lips and with a wide-open mouth, since it was a common problem with the previous tests with the Kivent game; while the 2nd and 3rd tests were meant to enable the development of our adaptation of the BEV evaluation algorithm used in regular spirometry.

We attempted to collect at least one successful recording of each type. At the end of each manoeuver, the healthcare professional registered its quality with the in-game form in a “yes”/”not sure”/”no” format, along 6 categories as previously shown on Figure 22. A total of 55 children took part in the data collection, producing a total of 400 audio samples. The distribution of those samples among the different tests is presented on Table 3.

Table 3 - Total amount of recordings obtained per test

Test Type	Count
Cold Air Test	167
Hot Air Test	88
Max Force Test	144

6.2 Algorithm Development

6.2.1 Minimal Effort

Throughout the data collection events, we noticed that it was common for children to not make a meaningful effort when performing the first test, producing a barely audible exhale. This led the healthcare professional to have to repeat the first test multiple times on average, while encouraging the child to exhale with greater force, which explains the highest number of attempts performed on what would be the easiest test, as can be seen on Table 3. However, it should be noted that this result isn’t entirely surprising either, as one of the consequences of performing these tests with mostly non-asthmatic children is that they have little to no experience in performing exhalation-related evaluations.

From observation during the tests, interferences were rare compared to sub-minimal effort issues. Looking at the labelled database after the fact, it can be observed that only 11 of all

manoeuvers (2.6%) were classified as having interferences. However, measuring the occurrences of minimal effort is not as trivial given there was no specific classification for it during our tests. Therefore, we used a conservative approach to count the samples that had a sub-minimal effort: by selecting all samples that were labelled with unsatisfactory initial, continuous, and finish criteria simultaneously. With this, we counted 38 manoeuvers (10.5%).

Due to having much fewer samples to develop and validate work on automatic interference detection compared to minimal effort evaluation, we chose to focus on the latter instead of the former.

The purpose of this criterion will be simply to reduce the amount of bad quality signals that will be passed on to the main processing pipeline, providing some assurance that the evaluated signal contains any section that could possibly represent a meaningful exhale.

6.2.1.1 Data Set

From the labelled database we selected the previously mentioned 38 manoeuvers as our negative cases, while the positive cases were selected as the ones having satisfactory initial, continuous and finish criteria simultaneously. We then randomly split each of these groups into two sub-groups, one for algorithm development purposes and the other for validation purposes, using a 70/30 split (Table 4). Given the low sample count particularly for the validation set, we re-shuffled the development and validation sets two more times to create a total of 3 development and validation set versions, to reduce overfitting issues.

Table 4- Data set distribution for minimal effort evaluation

	Positive Cases	**Negative Cases**
Total	158	38
Development Set	110	26
Validation Set	48	12

6.2.1.2 Design and Implementation

In designing the algorithm to evaluate this criterion, we attempted to define what a meaningful exhale is in two dimensions – time and energy.

As a starting point regarding minimum acceptable time, we looked in child spirometry literature for the lowest acceptable manoeuver duration. Covering the ages of 3 to 8 years old (Gochicoa-Rangel et al., 2013) (Pesant et al., 2007) (Piccioni et al., 2007) (Linares P. et al., 2007), it was commonly observed that the forced expiratory volume in the first 0.5 and 0.75 seconds of the manoeuvre ($FEV_{0.5}$ and $FEV_{0.75}$) are argued to be more useful in interpreting spirometry in this age group. Therefore, we defined the time threshold T_t that a given signal's envelope must stay above the noise level as 0.5 seconds, in order for it to have a chance of being considered meaningful.

As will be further demonstrated on the mouth aperture algorithm development, the possible frequency spectrum for exhalation signals is very wide and therefore not easily defined on the

frequency domain. Therefore, we turned to time-domain thresholding methods to determine the noise level on our samples.

We calculated a simple root mean square envelope of each audio sample of the development dataset (exemplified in Figure 23) and analysed their histograms (figure 24). The histograms shown are representative of the whole, as almost all samples follow the same skewed unimodal distribution (only 1 of the positive cases did not, but this was due to external interferences). Due to this, we implemented and used the triangle thresholding algorithm (Zack et al., 1977) given it was designed specifically for this situation.

On this algorithm a line is constructed between the maximum of the histogram on the left and the last value above 0, points A and B. The length of the line's normal to the histogram is then computed for each point between A and B. The level where the distance between the histogram and the line is maximal is then defined as the threshold value. An application of this algorithm is exemplified on one of the development samples on figures 25 and 26.

Having defined the noise threshold, the algorithm for minimal effort evaluation determines if the RMS envelope of the signal contains any contiguous section for a duration greater than T_t. If so, it then estimates the area of each of these sections through a simple summation.

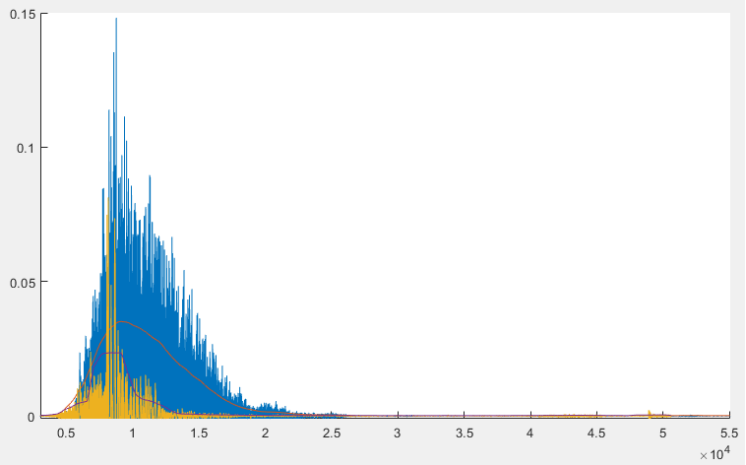

Figure 23 - In blue and yellow, the audio signals of a positive and a negative case from the development data set, respectively; in orange and red, their respective RMS envelopes

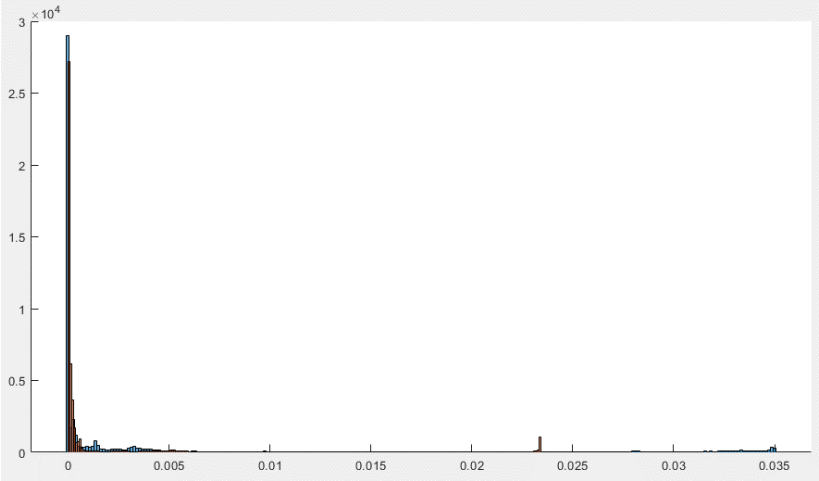

Figure 24 – In orange and blue, the RMS envelope's histogram of a positive and negative case from the development data set, respectively

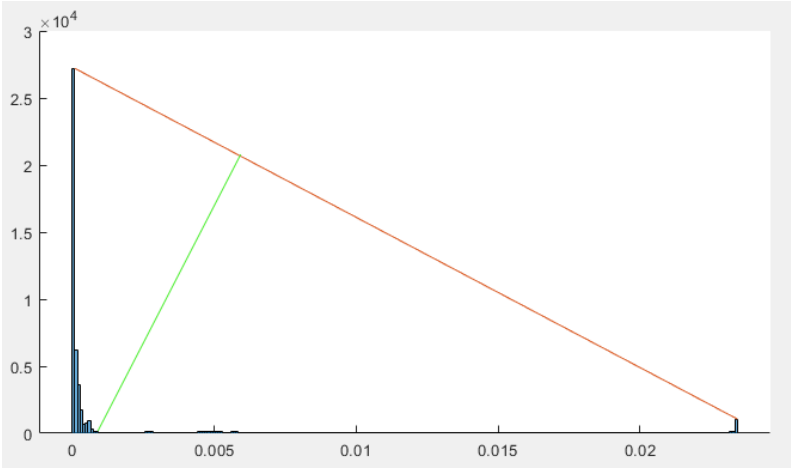

Figure 25 - Application of the triangle threshold algorithm on a histogram. The green line is the orange line's normal with maximal length between the orange line and the histogram

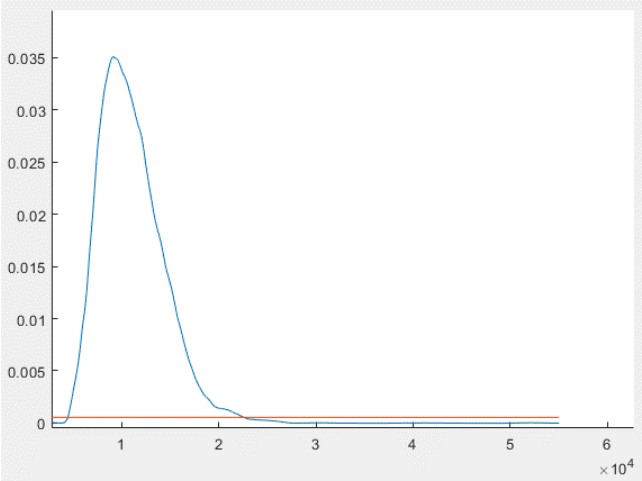

Figure 26 - In blue, the RMS envelope of the signal; in orange, the noise threshold level of the envelope according to the triangle algorithm;

And finally, if any of these integrals are above an area threshold T_a, we will consider the signal to contain an exhalation with minimal effort. To determine T_a we were unable to use a more empirical method as was used for the minimum duration's definition. Due to not having access to any absolute values of spirometry parameters, it was not possible to use well established points of reference presented in clinical spirometry literature. Therefore, we calculated the maximum area of sections above T_t for each sample on the development dataset. As an example, the area histograms for the negative and positive cases of the development dataset's first version are shown on Figure 27.

We then proceeded to pick the threshold that would maximise the following function:

$$Q = \sqrt{1.3 * TNR^2 + TPR^2}, TNR = \frac{TN}{TN + FP} \bigwedge TPR = \frac{TP}{TP + FN} \quad (1)$$

Where TN, TP, FN and FP represent the count of true negatives, true positives, false negatives and false positives, respectively, and TNR and TPR stand for true negative rate and true positive rate respectively.

The reason for the added weight to the true negative detection is due to the larger impact of false positives compared to false negatives in our context: a false positive would mean that we would be reinforcing bad form in the manoeuvre, while a false negative will only have the child repeat the manoeuvre. However, disregarding the true positive rate entirely could have a potentially discouraging effect to the child, so the added weight of 30% was assigned for the importance of TNR over TPR.

Having defined how the T_t and T_a thresholds will be obtained, this concludes the design stage of the minimal effort evaluation criterion.

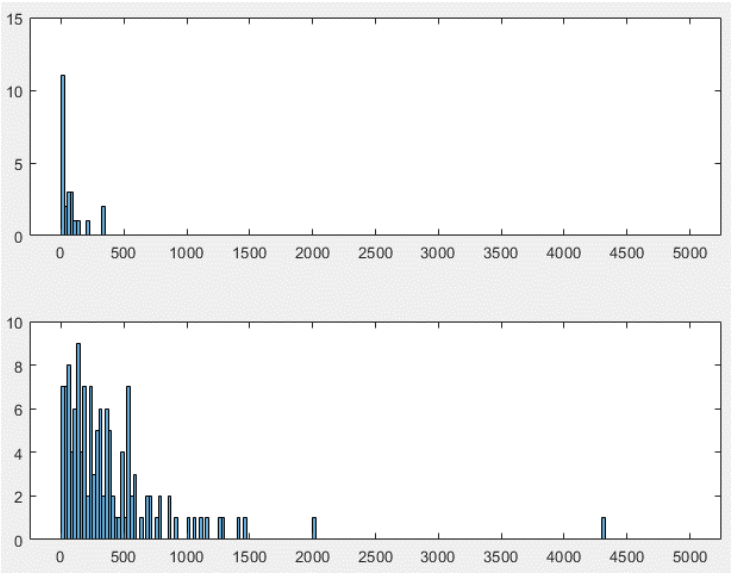

Figure 27 – Histograms showing the maximum area for contiguous sections above T_t, using the development dataset #1, for the negative (top) and positive cases (bottom).

6.2.1.3 Validation

In table 5 we can observe the performance of the minimal effort evaluation algorithm in all 3 versions of the dataset. The consistency of the development dataset's performance, allied with optimal thresholds around the same values, suggests that the significant performance disparity between the validation sets is possibly due to low sample counts.

In an attempt to better compare the algorithm's performance on the development and validation datasets, we used the Matthews correlation coefficient (MCC) (Matthews, 1975). We chose this classifier because it considers true and false positives and negatives with equal weight and allows the calculation of a single all-encompassing score, ranging between -1 and 1, where 1 and -1 indicate total agreement/disagreement between prediction and observation, while 0 indicates performance equal to random prediction. It is a commonly used as a performance metric in Bioinformatics, (Boughorbel et al., 2017) and can be determined by the following formula:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

Calculating the average MCC for the development and validation datasets, we obtain that the validation dataset scores 26% lower than the development dataset. Besides the fact that lower performance scores on validation datasets are not surprising, it is noteworthy that, already in this first attempt, we have managed to produce a classifier that is significantly better than randomly guessing the classification of a healthcare professional, for this specific criterion.

Table 5- TPR and TNR for each of the version of development and validation datasets, with the T_a obtained from the corresponding development dataset

Datasets	**T_a**	**Development Dataset**			**Validation Dataset**		
		TPR	TNR	MCC	TPR	TNR	MCC
Dataset Shuffle #1	150	75.38%	83.33%	0.589	54.17%	90.00%	0.473
Dataset Shuffle #2	125	79.23%	87.50%	0.670	75.00%	75.00%	0.500
Dataset Shuffle #3	100	81.68%	83.33%	0.650	84.21%	60.00%	0.456

However, there is a drawback to this method: it is device specific. Given part of it depends on the absolute measurement of the signal's energy (more precisely, a simplification thereof), and there are no guarantees as to the embedded microphone's sensitivities. This means that T_a must be tuned from device to device to ensure proper functioning. While this could be a significant limitation for a generic solution, it can be overcome in the specific context of Ar.Cade, as it is expected that the healthcare professional will perform an initial setup on the child's game for calibration purposes (among other purposes). Moreover, given the planned

medical back-office, it is feasible that during the healthcare professional's routine reviews of their patient's manoeuvre recordings they can perform their labelling, particularly during the first few weeks after the game is setup, feeding more data to the algorithm and enabling more precise classification.

The final quality analyser module will implement this algorithm with T_t at 0.5 seconds and the T_a threshold at 150, given our preference for higher TNR and the overall MCC for each of the thresholds being similar within the same dataset type.

6.2.2 Mouth Aperture

Since the project has the goal of not being dependent of any component external to the mobile device, the presence of a mouth piece to maintain adequate mouth aperture during microphone spirometry is not possible. Exhaling with pursed lips restricts airflow, and airflow restriction is a quality criterion defined in the ATS/ERS guidelines for clinical spirometry.

In the initial data collection, we noticed the difficulty of children in maintaining proper mouth aperture: in the hot air or maximal effort tests, 17 out of 55 children (30.9%) performed at least one manoeuvre with unsatisfactory mouth aperture after having already performed at least one successfully. With constant coaching of what proper mouth aperture looks like (often demonstrating multiple times per child), and even after children showing that they have learned it properly, a significant percentage still defaulted to exhaling with pursed lips as that was more natural to them. We would argue that without the presence and coaching of a healthcare professional this percentage would be higher. This shows the importance of the development of an algorithm for automated mouth aperture evaluation (along with proper feedback to the child).

6.2.2.1 Data Set

We selected the samples according to the positive or negative classification of the healthcare professional on the "appropriate mouth aperture" label, leaving out 5 samples classified as "not sure" as we have effectively no ground truth for them. Just as in the minimal effort evaluation data set, we used a 70-30 split for the development and validation sets, as can be seen on Table 6. On this data set however, not only do we have a significantly larger sample count, but the classes are also much more balanced.

Table 6– Data set distribution for mouth aperture evaluation

	Positive Cases	Negative Cases
Total	230	165
Development Set	161	115
Validation Set	69	50

6.2.2.2 Design and Implementation

To the best of our knowledge, there is no literature discussing the detection of mouth aperture via audio. Therefore, we analysed the development data set in both time and frequency domains looking for patterns that would allow to distinguish both. While there were no significant differences we could find on the time-domain, they were obvious on the frequency domain as can be seen on Figure 28. Specifically, the relative prevalence of lower frequency components over others on pursed lips exhalations.

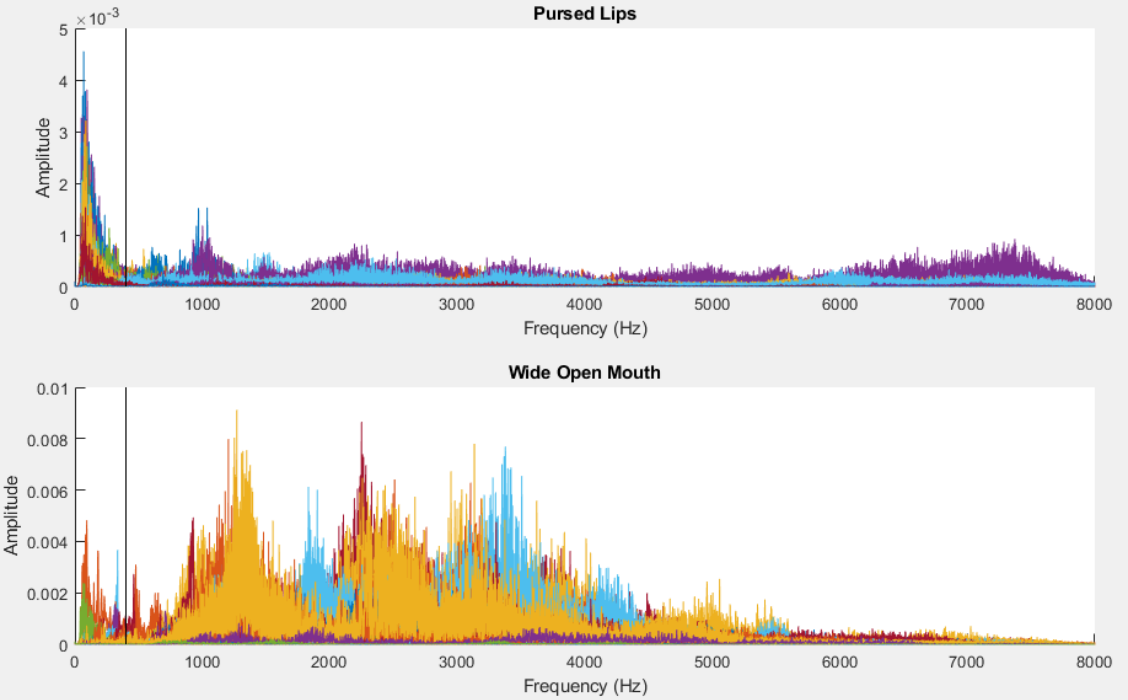

Figure 28 - Overlapped FFT's of all pursed lips (top) and wide open mouth exhalations (bottom). The black vertical line is placed at 400Hz.

We then began to develop a classifier based on the low frequency ratio, defined as the ratio between the FFT's summation below a frequency threshold T_f and the whole FFT's summation. Through visual inspection, we defined T_f at 400Hz as can be seen on the black vertical line on Figure 28, as it is a clear threshold on the pursed lips FFTs. A histogram of the resulting ratios for each sample on the development dataset is shown on Figure 29. We finished this classifier by defining a low frequency ratio threshold T_{rf} to distinguish pursed lips from open mouth exhalations, by choosing the T_{rf} that maximizes Equation 1.

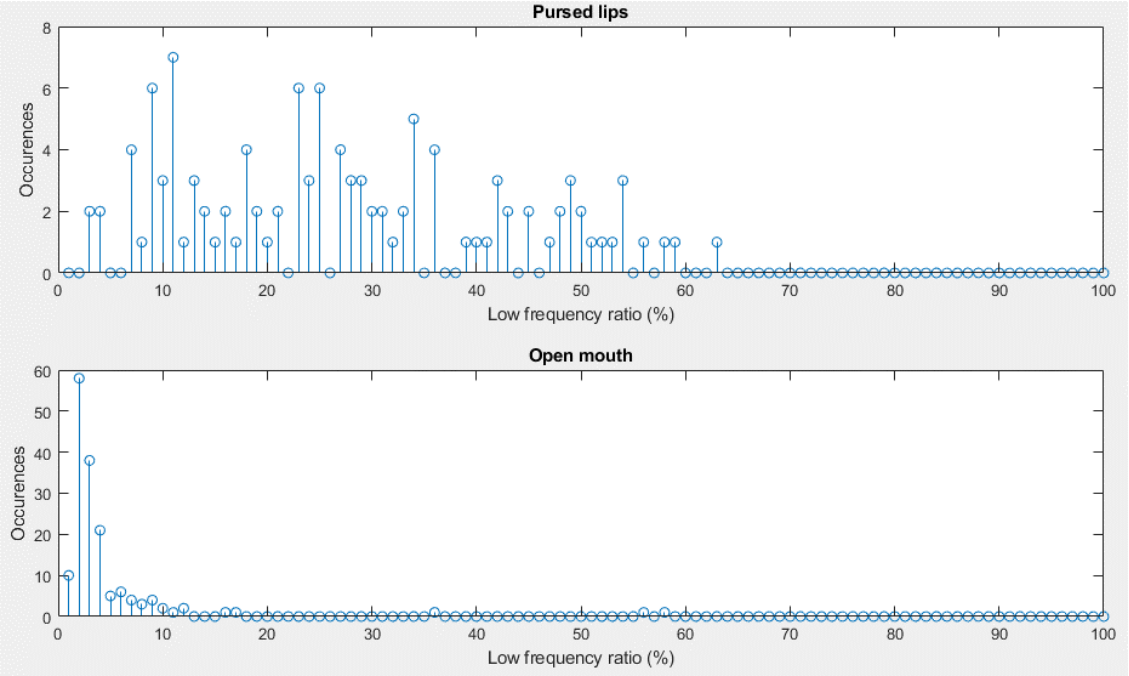

Figure 29 - Histogram of the low frequency ratios for pursed lips (top) and open mouth (bottom) samples of the development dataset

6.2.2.3 Validation

On Figure 30 we can observe the performance of different low frequency ratio thresholds evaluated on the development dataset. Using the Q formula, we obtained that the optimal threshold is 8%, which was then used on the validation dataset. These results are presented on Table 7.

As can be seen, the classifier not only shows a consistent performance between the development and validation datasets, but also the performance is strongly and positively correlated to the healthcare professional's labelling. While this may be a satisfactory classifier for this population, further testing should be performed not only on more children in general, but particularly on children of the lower half of our target age group, from 5 to 7-8 years old, to determine how well the observed pattern holds there.

Table 7 - TPR, TNR and MCC values for T_f = 400Hz and T_{rf} = 8% on the development and validation datasets

	TPR	TNR	MCC
Development set	92.04%	89.31%	0.8138
Validation set	93.75%	83.58%	0.7773

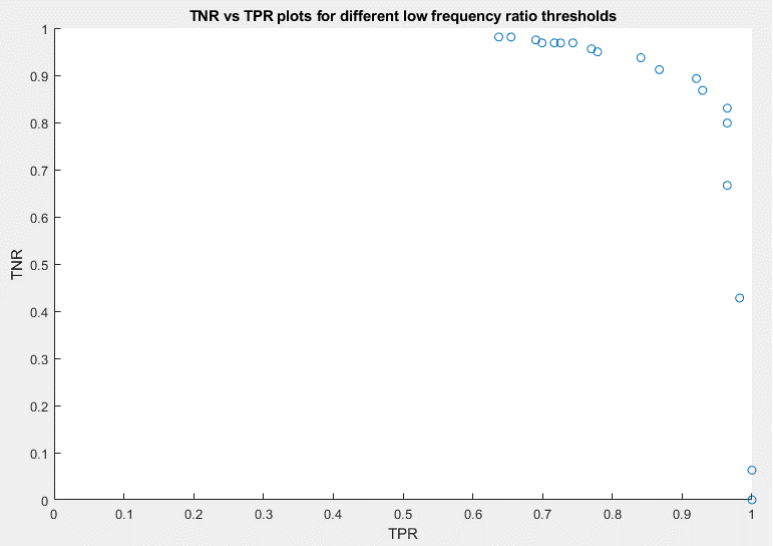

Figure 30 - TPR vs TNR plots for the different low frequency ratio threshold of the development dataset

85% of the negative cases on our dataset come from cold air tests, and those tests contain 58% of the cases we classified as below minimal effort. This is understandable given children had no prior experience playing the game, so worse performance is expected on these. However, the labelling of open mouth or pursed lips exhalation by the healthcare professional was based more on visual observation rather than audible, as even if the child did not exhale with any meaningful force it was possible to visualize the mouth's aperture. This in turn led to the low rate of "not sure" labels on the mouth aperture category (n=5, 1.2%). While in above minimal effort circumstances there were few discrepancies (i.e. the sound of a pursed lip exhalation matching the sound of a wide-open mouth exhalation or vice-versa), they are more likely in cases of sub minimal effort: regardless of the child's mouth aperture on exhale the signal will be mostly composed of the white noise captured by the microphone (ignoring cases of interference presence).

Given that a white noise signal has equal intensity at different frequencies, this means that the low frequency ratio will always tend to the frequency threshold coefficient

$$Frequency\ threshold\ coefficient = \frac{Frequency\ threshold}{FFT\ highest\ frequency} \quad (3)$$

Which in our case is $\frac{400\ Hz}{8000\ Hz} = 0.05 = 5\%$, lower than the proposed T_{rf}. This in turn implies that our algorithm tends to classify sub minimal efforts as wide-open mouth exhalations, while the healthcare professional classified 54% of sub minimal efforts as pursed lips.

To evaluate the actual impact of this potential source of false positive errors on our algorithm, we repeated our calculations after removing sub minimal efforts from the full dataset, once according to the healthcare professional's classification of sub minimal efforts, and once

according to the classification obtained by the minimal effort algorithm, with the AUC threshold set at 150. These results can be seen on Table 8.

Table 8 – Mouth aperture algorithm performance on three datasets: the original as described on 6.3.1; the original without sub-minimal effort exhalations as classified by the healthcare professional; and the original without sub-minimal effort exhalations as classified by our developed minimal effort evaluation algorithm.

Datasets	**Development Dataset**			**Validation Dataset**	
	TPR	TNR	T_{rf}	TPR	TNR
Original	92.04%	89.31%	8%	93.75%	83.58%
Healthcare professional	93.94%	90.48%	8%	92.86%	88.71%
Minimal effort algorithm (T_a = 150)	92.46%	89.92%	8%	93.12%	86.12%

As predicted, the TNR did increase with the removal of sub minimal efforts from the datasets, more so with the healthcare professional's classification than with our algorithm. Even so, this suggests that using the minimal effort algorithm as a precursor will increase the performance of the mouth aperture detection.

Another limitation with this algorithm is shown on Figure 31. This is an edge case we have encountered but cannot rule out as an outlier because on the 2 children where this was observed, it happened consistently. Due to lip moisture that was present when exhaling with pursed lips and with maximal effort, a slight whistling effect occurred that produced a perfectly plausible frequency profile for an open mouth exhalation, and indeed is considered as such by our algorithm. Considering the results already obtained with our classifier, this suggests that while it can be used as a first stage for mouth aperture evaluation, further stages should be developed. An approach worth exploring is by FET evaluation: as pursed lips restrict the airflow, one of the consequences is a falsely inflated duration of manoeuvre, which could be used to further detect improper mouth aperture.

Given the results presented in this classifier, for the Quality Analyzer we will implement it with frequency threshold T_f at 400Hz and low frequency ratio threshold T_{rf} at 8%.

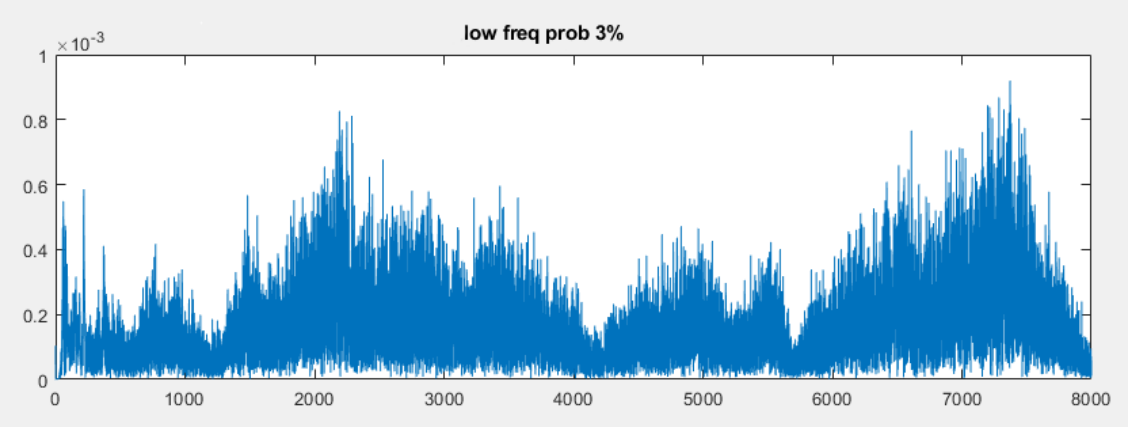

Figure 31 - FFT of a maximal effort pursed lips exhale with a slight whistling effect.

6.2.3 Back Extrapolated Volume

As presented in Chapter 3.3.1.1, official guidelines defined by the ATS (Miller et al., 2005) state that BEV should be lower than 150mL or 5% of the FVC, whichever is higher. As a first attempt to automatically determine if the manoeuvre's initial effort was acceptable or not, the official guidelines were used as much as possible, and we implemented an algorithm for BEV calculation according to ATS standards (Miller et al., 2005).

6.2.3.1 Data Set

Given that for this BEV evaluation we aren't building a new algorithm or aiming to define a new BEV threshold, but rather implementing the existing algorithm in clinical spirometry as closely as possible, we have no need for a development dataset like in the previous two algorithms. We will validate BEV thresholds presented in clinical spirometry literature against our entire dataset, concluding on their applicability (as well as on the algorithm as a whole) in the context of microphone spirometry.

Table 9 shows the used validation dataset, distinguishing hot air tests from maximal effort tests. We separated the cold and hot air tests from the maximal force test as the initial effort classification by the healthcare professional was more stringent for the latter.

Table 9 – Total amount of recordings obtained per test, and the distribution of acceptable BEV quality classification by the healthcare professional.

Validation datasets	Total	Yes	No
Dataset 1 - Cold and Hot Air Tests	87	156	95
Dataset 2 - Max Force Tests	141	108	33

6.2.3.2 Design and Implementation

Given that the BEV calculation is designed for the volume-time curve, for this algorithm we made use of the audio processing pipeline by (Teixeira et al., 2015), presented in Chapter 3.1.1, to extrapolate that curve from the audio signals. Using an 80ms window as suggested on the official guidelines for computerised approaches (Miller et al., 2005), we followed the standard procedure to determine the t_{n0} as described in 3.3.1.1.

Given that we do not have access to absolute values in our implementation of the processing pipeline, we chose to only use the relative criterion of BEV < 5% of FVC. FVC was calculated simply by performing a summation of the volume-time curve. We then compared the results obtained by this classifier with the healthcare professional's classification of the manoeuvres, acting as our ground truth.

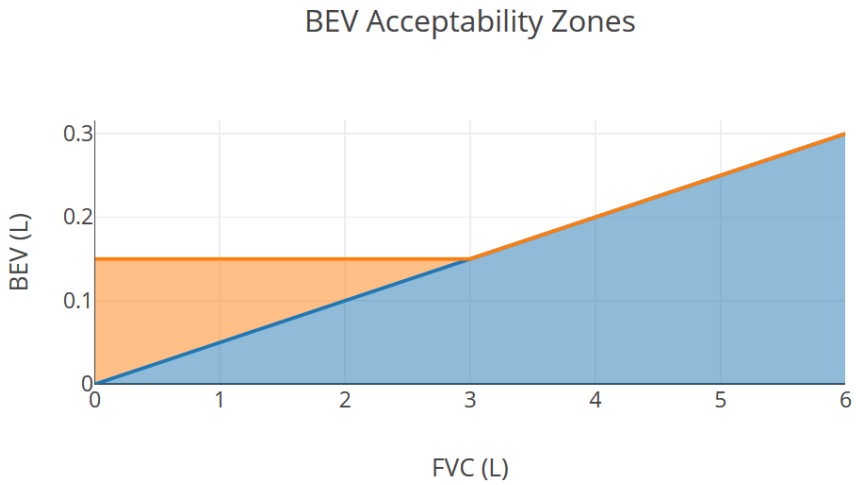

Figure 32 - In orange, the acceptable BEV ranges using the ATS guidelines; in blue, the ranges using the implemented algorithm (The ATS ranges overlap with the algorithm's ranges).

In figure 32 it is shown that for FVC values under 3 litres, the ATS guidelines are increasingly more lenient as the FVC decreases compared to just using the relative criteria implemented in the algorithm. To evaluate how relevant this issue is for our collected data, we used children's age specific reference equations for FVC developed by (Koopman et al., 2011). According to these equations, FVC varies with age, height and sex. We used the self-reported age and sex, while for height the World Health Organization's height-for-age charts were considered (de Onis et al., 2007) ("WHO | Height-for-age (5-19 years)," n.d.).

To illustrate the FVC boundaries of our targeted population the mean FVC, along with the lower and upper limit of normal (LLN and ULN) for a 5% cut-off, they are shown in tables 10 and 11.

Table 10 - FVC percentiles (5% cut-off) for 10-year-old children in the 99th height percentile.

	LLN	**Mean**	**ULN**
Male	2.52L	3.02L	3.65L
Female	2.35L	2.86L	3.41L

Table 11 - FVC percentiles (5% cut-off) for 5-year-old children in the 1st height percentile.

	LLN	Mean	ULN
Male	0.73L	0.91L	1.11L
Female	0.82L	1.00L	1.19L

This allows us to conclude that, for our target age group, we are more likely to be dealing with cases where the expected FVC is under 3L, making our algorithm more stringent than the official ATS guidelines. As an initial attempt to overcome this limitation, we defined an age adjusted T_{bev} threshold for each of the target ages. This threshold was calculated as thus:

$$T_{bev} = \frac{0.150ml}{FVC_i} \quad (4)$$

Where *i* represents a given child's age, and FVC_i is the mean expected FVC value for a given *i*, assuming the 50th height percentile of the WHO standards. The resulting BEV thresholds for males and females can be seen on table 12 and 13, respectively. By doing this, we are trying to have the acceptable BEV for a specific age match the static 150mL criterion defined by the ATS, in the context of a lack of access to absolute values.

The value found for 5-year-old males is in complete agreement with the findings of (Aurora 2004) in pre-schoolers' observed BEV/FVC ratio. In that work, a possible quality control cut-off of 12.5% is suggested, and therefore we will also evaluate its utility.

Table 12 – Age adjusted BEV thresholds for the 50th height percentile of each presented age in males.

Age	**5**	**6**	**7**	**8**	**9**	**10**
T_{bev} (%)	12.5	10.7	9.3	8.2	7.0	6.4

Table 13 – Age adjusted BEV thresholds for the 50th height percentile of each presented age in females.

Age	**5**	**6**	**7**	**8**	**9**	**10**
T_{bev} (%)	12.9	11.1	9.8	8.6	7.5	6.6

6.2.3.3 Validation

Tables 14 to 16 show the performance of the algorithm for 5% FVC, 12.5% FVC and age adjusted BEV thresholds. At a first glance the most notable feature is the MCC being close to 0 on all instances, showing an almost non-existent relationship between the algorithm's predictions and the healthcare professional's classifications.

For the 5% threshold, the low TPR indicates that the algorithm is too strict with respect to the healthcare professional. Nevertheless, given that the only information available to the healthcare professional to produce an evaluation was the visual observation of the child, and not an objective measurement obtained by a spirometer as happens on clinical spirometry, the chance of having some mislabelled ground truth is not negligible.

Given the already explained influence of age, sex and height on the FVC, it is worth noting that the presented results are based off a sample database that is heavily biased towards the higher end of the age spectrum. Given the positive correlation between age and height to expected FVC values, this means that it would be reasonable to expect our results to be worse with a more balanced database in terms of age distribution, at least for any static BEV threshold.

Table 14 – Performance for the BEV algorithm for the threshold 5% FVC.

	T_{bev} = 5%		
	TPR	TNR	MCC
Dataset 1	36.4%	71.5%	0.084
Dataset 2	41.7%	65.8%	-0.077

Table 15– Performance for the BEV algorithm for the threshold 12.5% FVC.

	T_{bev} = 12.5%		
	TPR	TNR	MCC
Dataset 1	85.7%	9.6%	-0.072
Dataset 2	88.8%	19.2%	0.111

Table 16– Performance for the BEV algorithm for the age adjusted threshold.

	Age adjusted T_{bev}		
	TPR	TNR	MCC
Dataset 1	66.7%	38.0%	0.049
Dataset 2	71.1%	35.5%	0.071

According to (Koopman et al., 2011) the direct correlation between FVC and age is small compared to FVC's correlation with height, as can be seen on Figure 33. Given this discrepancy and the fact that we assumed the average height for each age shows a limitation in our age adjusted BEV thresholds, as the height for a child varies around 20% from the 3rd percentile to the 97th, at any given age in our target group, according to WHO standards. For future work, children's height should be collected and used to evaluate the impact on BEV estimation.

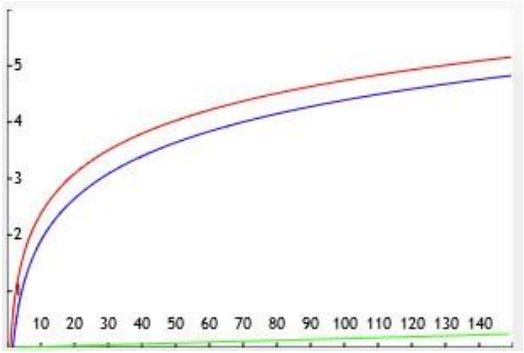

Figure 33 – Plot illustrating the importance of the predictors for FVC estimation as they vary with age or height. The red and blue lines represent the impact of children's height, measured in cm, for females and males respectively (target group bounded between 60 and 120 months of age)

Despite the limitations described above, it is important to note one final point. On Table 17 we can see the number of children that managed at least one positive detection with the current algorithm, set at different thresholds, considering the maximal force tests where they received specific coaching to that effect. Even for our worst performing threshold we obtained reasonably similar results to (Tomalak et al., 2008), where 80.4% of 117 children between the ages of 4 and 10 years old were able to pass the ATS standard for BEV acceptability, using clinical spirometers.

Table 17 - Number of children capable of performing at least one maximal force test with acceptable BEV for different thresholds, in absolute and relative units respectively.

Criterion	Children with at least one positive classification
Healthcare professional	54 (98%)
5% FVC	46 (84%)
12.5% FVC	55 (100%)
Age adjusted	52 (95%)

We defined that the 5% threshold would be the one to implement, given all thresholds have roughly the same overall performance compared to the healthcare professional (as evidenced by the MCC), but this one maximized the TNR while still attributing at least one positive classification to most children.

6.3 Feedback Messages Development

For the feedback system, we chose to establish a hierarchy in terms of quality criteria:

1. If minimal effort classification is positive, evaluate mouth aperture;
2. If mouth aperture classification is positive, evaluate BEV;

This was done with the following reasoning: we do not want to split the children's attention or overwhelm them with more than one instruction for improvement per attempt. Given we are only going to deliver one instruction directed for a single criterion per attempt, we would be wasting time and decreasing performance if we didn't stop evaluating criteria after the first negative.

As mentioned in Chapter 6.2.2.3, there is a default bias of the algorithm to classify a manoeuvre with positive mouth aperture in sub-minimal effort situations, and that the mouth aperture's classification performance increases if we filter out the sub-minimal effort manoeuvres. Therefore, we put minimal effort evaluation before mouth aperture evaluation.

When it comes to BEV evaluation, two arguments were used to define its hierarchy: first, we are simply not interested in evaluating a clinical parameter of a manoeuvre with poor form, because the number is meaningless in that case. If the manoeuvre has sub-minimal effort, then

we already know that BEV evaluation cannot be positive, and if the child is exhaling with pursed lips then they are restricting their maximal flow rate, which has a direct impact on BEV.

Secondly, while the other criteria can be evaluated almost directly from raw audio, the evaluation of this criterion requires the calculation of the Volume-Time curve, which in turn requires the use of the processing pipeline by (Teixeira et al., 2015). This requires much more expensive calculations and performing them without need may needlessly hinder gameplay.

Each quality classification was associated with a certain star count to be attributed to the player at the end of level, as follows:

0. Manoeuvre classified as sub-minimal effort
1. Manoeuvre classified as above minimal effort but with pursed lips
2. Manoeuvre classified as above minimal effort with appropriate mouth aperture, but with failed BEV evaluation
3. Manoeuvre fully classified as correct (within the context of the implemented criteria)

Having defined this, the last missing step for the completion of the feedback system is the actual definition of the audio feedback to be provided to the player, depending on the result of the quality analysis.

This audio feedback was designed based on observations from the initial data collection event. A simple and short script (annex C) was written with the same type of colloquial language used by the healthcare professional during that data collection for coaching, being recorded for in-game playback. Four audio files were recorded in total, one for each criterion failure and one for full success.

6.3.1 Field Validation

To validate this feedback system, we performed a second data collection event, whose protocol is presented on annex D. This time we specifically planned no interventions or coaching by the healthcare professional, nor before or during the playtest. Players used headphones, and the game device was positioned such that the healthcare professional could not see or hear any feedback from the game, removing the possibility for having their judgement in classification influenced by the game and the quality analyser.

This data collection event was geared only for the children that had already performed the first test. While it would be interesting and important we would validate the algorithms on a different population, we are not trying to perform validations of the algorithm. Instead, we are now trying to answer the question “can we remind children with previous training how to perform the manoeuvre without human intervention?”. The previous training factor is important because that is the planned use case of Ar.Cade: children are expected to have at

least one consultation with the relevant healthcare professional where their game will be setup.

At the end of each manoeuvre, children received the appropriate audio feedback as determined by the game. However, during the event we noticed that no child ever received a negative classification on the BEV criterion, which was completely unexpected as not only is it our worst performing algorithm, but it was specifically biased towards false negatives instead of false positives as previously explained. Upon further code inspection after the fact, we noticed a typo on the configured BEV threshold: 0.5 instead of 0.05, that is 50% instead of the desired 5%. This means that we were thus unable to validate the specific feedback associated with BEV criterion failure, as due to time constraints it was not possible to return a second time for this second phase data collection.

A total of 20 children participated, and it is also worth mentioning that the timespan between the initial data collection events and this second phase was over a month for all children involved, some going up to one month and three weeks.

On table 18 we can see the significance of this fact: most children did not remember how to properly perform the FS manoeuvre adequately: they either exhaled with not enough effort, or they exhaled with pursed lips aside from 2 children.

Table 19 shows that, at most, a child was stuck on the same sub 3-star classification during 4 attempts, with the majority progressing after listening to a given feedback just once. The difference between total attempt count on table 17 and 18 is due to almost half the children never receiving a 0-star ranking.

Table 18 – Classifications received by children on their 1st attempt.

Manoeuvre Classification	**Count**
0	8
1	10
3	2

Table 19- Number of attempts each child required to progress from a 0 and a 1-star ranking

Attempts until improvement	0-star ranking	1-star ranking
1	4	10
2	2	2
3	1	2
4	0	1
5	0	0

Table 20- Number of attempts each child required until the first 3-star rating

Attempts until max rank	Count
1	2
2	6
3	7
4	3
5	2

The most interesting result however is that on table 20. Every child took less than 6 attempts to reach a 3-star classification, and the significance of this is amplified by the timespan between their last coaching and overall interaction with the game. More interestingly, the performance of this system falls within the guidelines of regular spirometry as per the ATS, in the sense that a maximum quality manoeuvre was attained within 8 attempts (Miller et al., 2005). All children exhibited interest in playing the game and did not grow frustrated at any point.

However, several limitations must be noted. First, this success within 8 attempts is only a success for a subset of quality criteria. There are more that require evaluation and yet were not even implemented or studied, aside from the BEV criterion which was implemented but not correctly configured. With its correct presence, we would expect the attempt count until 3 stars to increase quite possibly above 8 attempts given the lower performance of the algorithm and its bias towards false negatives.

The low number of children involved in this validation study, along with the restricted age group of 8 to 10 years old, limits our ability to extrapolate these results for the general target population of 5 to 10 years old. This not only applies in terms of 1) the current algorithm applicability for the general target population, where we need to verify how well the detected patterns hold, but also 2) in terms of the audio feedback's usability, provided given the lower cognitive capabilities of a 5-6 year old compared to the 8-10 year olds, generally speaking.

Also, we expect over fitting to have happened in terms of quality analysis algorithm performance, as we performed this field validation on a subset of the same population that was used for the algorithm development. While this is likely to have inflated the performance of the quality analyser, it was convenient in the sense that it helped isolate the performance of the audio feedback, which is what we were trying to measure in this field validation study.

This then suggests that assuming a quality analyser with good performance, it is possible to remind previously trained children how to perform good quality forced spirometry using an automated feedback system.

7 Conclusion

We developed a solution as an initial attempt at an automatic manoeuvre quality classifier, with the microphone as the input sensor. This solution was implemented in a mini-game to be integrated on the Ar.Cade project. It is capable of reminding previously trained children how to correctly perform a manoeuvre for an important subset of microphone spirometry quality criteria.

To achieve this, we performed two stages of data collection events in a controlled setting, the first stage led by a healthcare professional to enable the development of the quality classifier, and the second to perform its validation without the intervention of the healthcare professional.

The performance of the quality criteria analyser was overall satisfactory, however less so for one of the three criteria. We obtained a Matthew's Correlation Coefficient of 0.5, 0.77 and 0.084 on the validation datasets for minimal effort, mouth aperture and initial effort evaluation algorithms respectively. In terms of the results obtained from the feedback system, they were satisfactory as every child was able to obtain maximum classification relying only on the feedback provided through the game, without the healthcare professional having to intervene and before the maximum of 8 attempts were reached, as defined in current literature.

Our results reinforce the idea that microphone spirometry is a promising and viable field of study. However, studies with larger samples of children encompassing the entirety of the target age range are needed to further improve the system, and to ascertain the applicability of these results to the general population. To improve the robustness of the quality detection algorithms, the data collections should be cross-evaluated by more than one healthcare professional. These data collections themselves should include multiple devices to enable better understanding of how different microphone configurations impact the resulting signals.

References

Altaxo [WWW Document], 2015. URL https://github.com/Altaxo/Altaxo (accessed 7.5.18).

Andringa, T., 2014. Farseer-Physics: A Git Mirror of Farseer Physics a C# port of Box2D.

Arets, H.G.M., Brackel, H.J.L., Van Der Ent, C.K., 2001. Forced expiratory manoeuvres in children: do they meet ATS and ERS criteria for spirometry? 655–660.

Belliveau, P., Griffin, A., Somermeyer, S., 2004. The PDMA ToolBook 1 for New Product Development, Product Development and Management ToolBooks. Wiley.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE 12, e0177678. https://doi.org/10.1371/journal.pone.0177678

Chelaru, V., 2016. FlatRedBall – 2D Game Engine [WWW Document]. URL http://flatredball.com/ (accessed 7.4.18).

Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper [WWW Document], 2017. . Cisco. URL https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html (accessed 2.25.18).

Costa, A.M.P., 2016. Analysis of Pulmonary Airflow on a Smartphone Application.

Culver, B.H., Graham, B.L., Coates, A.L., Wanger, J., Berry, C.E., Clarke, P.K., Hallstrand, T.S., Hankinson, J.L., Kaminsky, D.A., MacIntyre, N.R., McCormack, M.C., Rosenfeld, M., Stanojevic, S., Weiner, D.J., ATS Committee on Proficiency Standards for Pulmonary Function Laboratories, 2017. Recommendations for a Standardized Pulmonary Function Report. An Official American Thoracic Society Technical Statement. Am. J. Respir. Crit. Care Med. 196, 1463–1472. https://doi.org/10.1164/rccm.201710-1981ST

de Onis, M., Onyango, A.W., Borghi, E., Siyam, A., Nishida, C., Siekmann, J., 2007. Development of a WHO growth reference for school-aged children and adolescents. Bull. World Health Organ. 85, 660–667.

Derom, E., van Weel, C., Liistro, G., Buffels, J., Schermer, T., Lammers, E., Wouters, E., Decramer, M., 2008. Primary care spirometry. Eur. Respir. J. 31, 197–203. https://doi.org/10.1183/09031936.00066607

Dewulf, K., 2013. Sustainable Product Innovation: The Importance of the Front- End Stage in the Innovation Process. Adv. Ind. Des. Eng. 139–166. https://doi.org/10.5772/52461

FIRS, 2017. The Global Impact of Respiratory Disease - Second Edition [WWW Document]. URL https://www.firsnet.org/publications/the-global-impact-of-respiratory-disease (accessed 2.13.18).

Gochicoa-Rangel, L., Vargas-Domínguez, C., García-Mujica, M.E., Bautista-Bernal, A., Salas-Escamilla, I., Pérez-Padilla, R., Torre-Bouscoulet, L., 2013. Quality of spirometry in 5-to-8-year-old children. Pediatr. Pulmonol. 48, 1231–1236. https://doi.org/10.1002/ppul.22765

Koopman, M., Zanen, P., Kruitwagen, C.L.J.J., van der Ent, C.K., Arets, H.G.M., 2011. Reference values for paediatric pulmonary function testing: The Utrecht dataset. Respir. Med. 105, 15–23. https://doi.org/10.1016/j.rmed.2010.07.020

Linares P., M., Meyer P., R., Cox F., P.P., Contreras E., I., 2007. Utility of FEV 0.5 versus FEV 1 in interpreting spirometry in preschool children with asthma. Rev. Chil. Enfermedades Respir. 23, 167–172.

Liu, X., 2013. mCOPD: Mobile Phone Based Lung Function Diagnosis and Exercise System for COPD. UCLA.

Luo, A.Z., Whitmire, E., Stout, J.W., Martenson, D., Patel, S., 2017. Automatic characterization of user errors in spirometry, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Presented at the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4239–4242. https://doi.org/10.1109/EMBC.2017.8037792

Math.NET Numerics [WWW Document], 2002. URL https://numerics.mathdotnet.com/ (accessed 7.5.18).

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta BBA - Protein Struct. 405, 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Melia, U., Burgos, F., Vallverdú, M., Velickovski, F., Lluch-Ariet, M., Roca, J., Caminal, P., 2014. Algorithm for Automatic Forced Spirometry Quality Assessment: Technological Developments. PLOS ONE 9, e116238. https://doi.org/10.1371/journal.pone.0116238

Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Crapo, R., Enright, P., Grinten, C.P.M. van der, Gustafsson, P., Jensen, R., Johnson, D.C., MacIntyre, N., McKay, R., Navajas, D., Pedersen, O.F., Pellegrino, R., Viegi, G., Wanger, J., 2005. Standardisation of spirometry. Eur. Respir. J. 26, 319–338. https://doi.org/10.1183/09031936.05.00034805

Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Enright, P., van der Grinten, C., Gustafsson, P., Jensen, R., Macintyre, N., McKay, R.T., Pedersen, O.F., Pellegrino, R., Viegi, G., Wanger, J., 2010. Standardisation of lung function testing: the authors' replies to readers' comments. Eur. Respir. J. 36, 1496–1498. https://doi.org/10.1183/09031936.00130010

MIR, 2014. MIR Smart One: Peak Flow and FEV1 for iOs and Android Smart Phones [WWW Document]. URL https://www.spirometry.com/Eng/Products/smartone.asp (accessed 7.4.18).

Moger, A., Holton, K., Hill, S., Kearney, M., Winter, R., 2013. Quality Assured Diagnostic Spirometry(QADS) - Performance And Competence, in: B45. CHRONIC OBSTRUCTIVE PULMONARY DISEASE: DIAGNOSIS AND EVALUATION, American Thoracic Society International Conference Abstracts. American Thoracic Society, pp. A2836–A2836. https://doi.org/10.1164/ajrccm-conference.2013.187.1_MeetingAbstracts.A2836

Müller-Brandes, C., Krämer, U., Gappa, M., Seitner-Sorge, G., Hüls, A., Von, B., Hoffmann, B., Schuster, A., Illi, S., Wisbauer, M., Berdel, D., 2014. LUNOKID: Can numerical American Thoracic Society/European Respiratory Society quality criteria replace visual inspection of spirometry? Eur. Respir. J. 43, 1347–1356. https://doi.org/10.1183/09031936.00058813

Pesant, C., Santschi, M., Praud, J.-P., Geoffroy, M., Niyonsenga, T., Vlachos-Mayer, H., 2007. Spirometric pulmonary function in 3- to 5-year-old children. Pediatr. Pulmonol. 42, 263–271. https://doi.org/10.1002/ppul.20564

Piccioni, P., Borraccino, A., Forneris, M.P., Migliore, E., Carena, C., Bignamini, E., Fassio, S., Cordola, G., Arossa, W., Bugiani, M., 2007. Reference values of Forced Expiratory Volumes and pulmonary flows in 3–6 year children: a cross-sectional study. Respir. Res. 8, 14. https://doi.org/10.1186/1465-9921-8-14

Salas, T., Rubies, C., Gallego, C., Muñoz, P., Burgos, F., Escarrabill, J., 2011. Technical Requirements of Spirometers in the Strategy for Guaranteeing the Access to Quality Spirometry. Arch. Bronconeumol. Engl. Ed. 47, 466–469. https://doi.org/10.1016/j.arbr.2011.06.006

Teixeira, J.F., Teixeira, L.F., Fonseca, J., Jacinto, T., 2015. Automatic Analysis of Lung Function Based on Smartphone Recordings, in: Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science. Presented at the International Joint Conference on Biomedical Engineering Systems and Technologies, Springer, Cham, pp. 390–402. https://doi.org/10.1007/978-3-319-27707-3_24

Tinschert, P., Jakob, R., Barata, F., Kramer, J.-N., Kowatsch, T., 2017. The Potential of Mobile Apps for Improving Asthma Self-Management: A Review of Publicly Available and Well-Adopted Asthma Apps. JMIR MHealth UHealth 5, e113. https://doi.org/10.2196/mhealth.7177

Tomalak, W., Radliński, J., Latawiec, W., 2008. Quality of spirometric measurements in children younger than 10 years of age in the light of the recommendations. Pneumonol. Alergol. Pol. 76, 421–425.

Velickovski, F., Ceccaroni, L., Marti, R., Burgos, F., Gistau, C., Alsina-Restoy, X., Roca, J., 2018. Automated Spirometry Quality Assurance: Supervised Learning From Multiple Experts. IEEE J. Biomed. Health Inform. 22, 276–284. https://doi.org/10.1109/JBHI.2017.2713988

Walters, J.A., Walters, E.H., Nelson, M., Robinson, A., Scott, J., Turner, P., Wood-Baker, R., 2011. Factors associated with misdiagnosis of COPD in primary care. Prim. Care Respir. J. J. Gen. Pract. Airw. Group 20, 396–402. https://doi.org/10.4104/pcrj.2011.00039

WHO | Height-for-age (5-19 years) [WWW Document], n.d. . WHO. URL http://www.who.int/growthref/who2007_height_for_age/en/ (accessed 4.3.18).

World Health Organization, n.d. WHO | 10 facts on asthma [WWW Document]. WHO. URL http://www.who.int/features/factfiles/asthma/en/ (accessed 2.25.18).

Zack, G.W., Rogers, W.E., Latt, S.A., 1977. Automatic measurement of sister chromatid exchange frequency. J. Histochem. Cytochem. 25, 741–753. https://doi.org/10.1177/25.7.70454

Zubaydi, F.K., 2016. A mobile Based Platform for Monitoring Respiratory Diseases (Thesis).

Annex A – Permission Request for Data Collection

Exmoº Encarregado(a) de Educação,

Vimos pedir-lhe a autorização para que o seu educando faça parte de uma colheita de dados do seu sopro, para efeitos de desenvolvimento de um jogo destinado a crianças asmáticas.

Este jogo que estamos a desenvolver tem como objetivo permitir a monitorização da asma fora do consultório, usando apenas os sensores já presentes em smartphones e tablets, em especial o microfone. Analisando o som produzido pela respiração da criança tencionamos extrair informação útil, quer para os pais quer para os profissionais de saúde, sobre a evolução da asma na criança. Somos uma *spinoff* do CINTESIS, um centro de investigação da Faculdade de Medicina do Porto, chamada MEDIDA e ligada à investigação e desenvolvimento na área da medicina.

Na fase atual do projeto precisamos de recolher amostras sonoras de crianças desta faixa etária a soprarem, independentemente de terem asma ou não.

A sessão será realizada nas instalações do Externato, com uma duração aproximada de 15 minutos por criança. Será guiada por um profissional de saúde que irá interagir com a criança, explicar como interagir com o jogo no *smart device*, e como é suposto realizar a respiração. As gravações serão feitas pelo próprio dispositivo do jogo no momento em que é esperado que a criança sopre, bem como noutros *smart devices* auxiliares em simultâneo. Nestas sessões estará também presente o engenheiro informático responsável pelo desenvolvimento da aplicação, que irá simplesmente observar a forma como a criança interage com o jogo, avaliando questões relacionadas com a usabilidade do mesmo.

Numa segunda fase, após a análise dos dados recolhidos, gostaríamos de repetir a atividade com o seu educando para validação dos resultados.

Oferecemos um rastreamento gratuito da capacidade respiratória do seu educando no final da atividade, a ser realizado pelo profissional de saúde.

Para mais informações ou esclarecimentos pode contactar Bernardo Pinho, o engenheiro responsável por organizar estas atividades por email bernardopinho@gmail.com.

Ficar-lhe-íamos gratos se permitir que o seu educando faça parte desta recolha de dados, imprimindo e preenchendo nos espaços em branco o pedido de autorização na página seguinte.

Eu, _____ (Encarregado de Educação), autorizo que o meu/minha educando(a) _____, faça parte da atividade de recolha de dados sonoros respiratórios para efeitos de investigação e desenvolvimento.

Assinatura: _____

Data: ___/___/_____

Annex B – First Data Collection Protocol

Pretende-se recolher dados para estudo da qualidade do som respiratório adquirido durante a manobra de expiração forçada executada em crianças no 1º ciclo do ensino básico (entre 5 e os 10 anos). O profissional de saúde é responsável por dirigir todo este processo.

1. Iniciar o jogo no dispositivo móvel
2. Introduzir os dados do participante.
3. Mostrar a criança ao Fred, explicar o intuito do jogo - soprar para destruir os blocos
4. Demonstrar uma vez com o jogo, mostrando à criança como é suposto interagir com ele.
5. Pedir à criança para ela jogar, com inspiração rápida e completa (encher o peito de ar) soprando da forma mais prolongada possível:
    a. Pelo menos uma manobra em que sopre ar frio (lábios semicerrados) Sugestão soprar uma vela
    b. Pelo menos uma manobra em que sopre ar quente (boa abertura da boca) Sugestão embaciar um vidro/espelho.
6. Registar no final de cada tentativa a qualidade da manobra usando a interface do jogo.
7. Explicar à criança como fazer uma manobra de expiração forçada.
8. Pede à criança para ela jogar:
    a. Pelo menos uma manobra com qualidade suficiente.
9. Registar no final de cada tentativa a qualidade da manobra, usando a interface do jogo.

Annex C – Feedback System Phrases

Table 21 - The feedback messages associated with each ranking

Ranking	Feedback Message
0	O dragão não te conseguiu ouvir, precisa que sopres com mais força para o motivares, senão o projétil não chega lá!
1	Boa! O dragão já conseguiu ouvir o teu sopro! Mas agora ele precisa que tu abras mais a boca, e que sopres ar quente para o projétil não arrefecer!
2	Espetacular! Com o teu sopro quente conseguiste ajudar o dragão a soprar uma bola de fogo! Só te falta soprares com mais força logo no início!
3	Brutal! Conseguiste ajudar o dragão a lançar o projétil mais forte de todos! Queres tentar de novo?

Annex D – Second Data Collection Protocol

Pretende-se recolher dados para estudo da qualidade da aprendizagem das crianças com base no *feedback* dado pelo jogo durante a manobra de expiração forçada, executada em crianças no 1º ciclo do ensino básico (entre 5 e os 10 anos). O profissional de saúde não deve poder observar nem escutar o feedback dado pelo jogo ao participante, de modo a não ser influenciado pelo classificador na sua avaliação.

1. Iniciar o jogo no dispositivo móvel
2. Perguntar à criança se se lembra das premissas básicas do jogo – do clicar no botão para começar e do imitar o Fred no momento de inspiração e de expiração.
    a. Em caso afirmativo, prosseguir para o ponto 3.
    b. Em caso negativo, relembrar as premissas sem dar instruções quanto à forma de expiração adequada. Após verificação da compreensão da criança, avançar para o ponto 3.
3. Pede à criança para ela colocar os auriculares.
4. Pede à criança para ela jogar, até que uma das seguintes condições se verifique:

a. A criança não querer continuar mais;
b. A criança atingir 3 manobras com sucesso consecutivas;
c. A criança executar 8 tentativas.

5. O profissional de saúde regista no final de cada tentativa a qualidade da manobra usando a interface do jogo.