



TECNOLOGÍA
en marcha

Quarterly journal
November 2022
Volume 35
ISSN-E 2215-3241

Special issue
International Work Conference
on Bioinspired Intelligence



TEC | Tecnológico
de Costa Rica

Publicación y directorio en catálogos



redalyc.org



DOAJ

Comisión Editorial

Felipe Abarca Fedullo. Director.
Editorial Tecnológica de Costa Rica

Juan Antonio Aguilar Garib
Facultad de Ingeniería Mecánica y Eléctrica
Universidad Autónoma de Nuevo León.
México

Carlos Andrés Arredondo Orozco
Facultad de Ingenierías
Universidad de Medellín. Colombia

Lars Köhler
Experimenteller Botanischer Garten
Georg-August-Universität Göttingen.
Alemania

Jorge Solano Jiménez
Instituto Costarricense del Cemento
y del Concreto

Edición técnica

Alexa Ramírez Vega

Revisión filológica

Esperanza Buitrago Poveda

Diseño gráfico

Felipe Abarca Fedullo

Diagramación

Leila Calderón Gómez

Diseño de cubierta

Ariana Sanabria García

Imagen de cubierta

Kevin Arias-Ceciliano

Datos de catalogación en publicación

Tecnología en Marcha / Editorial Tecnológica de Costa Rica. - Vol. 35, special issue. International Work Conference on Bioinspired Intelligence. November, 2022– Trimestral ISSN-E 2215-3241

1. Ciencia y Tecnología –
Publicaciones periódicas CDD:600



Apdo 159-7050 Cartago, Costa Rica
Tel.:(506) 2550-2297, 2550-2618
Correo electrónico: editorial@itcr.ac.cr
Web: <https://www.tec.ac.cr/editorial>
https://revistas.tec.ac.cr/tec_marcha



La Editorial Tecnológica de Costa Rica es una dependencia especializada del Instituto Tecnológico de Costa Rica. Desde su creación, en 1978, se ha dedicado a la edición y publicación de obras en ciencia y tecnología. Las obras que se han editado abarcan distintos ámbitos respondiendo a la orientación general de la Institución.

Hasta el momento se han editado obras que abarcan distintos campos del conocimiento científico-tecnológico y han constituido aportes para los diferentes sectores de la comunidad nacional e internacional.

La principal motivación de la Editorial es recoger y difundir los conocimientos relevantes en ciencia y tecnología, llevándolos a los sectores de la comunidad que los requieren.

La revista *Tecnología en Marcha* es publicada por la Editorial Tecnológica de Costa Rica, con periodicidad trimestral. Su principal temática es la difusión de resultados de investigación en áreas de Ingeniería. El contenido de la revista está dirigido a investigadores, especialistas, docentes y estudiantes universitarios de todo el mundo.

Publicación y directorio en catálogos





TECNOLOGÍA *en marcha*

Contenidos

Presentation special issue

Presentación número especial

Melvin Ramírez-Bogantes, Jose Luis Vásquez-Vásquez, Carlos M. Travieso-González 3

The impact of social media messages on Parkinson's disease treatment: detecting genuine sentiment in patient notes

El impacto de los mensajes de las redes sociales en el tratamiento de la enfermedad de Parkinson: detección de sentimientos genuinos en las notas de los pacientes

Hanane Grissette, El Habib Nfaoui..... 5

Automated adenocarcinoma lung cancer tissue images segmentation based on clustering

Segmentación automatizada de imágenes de tejido de cáncer de pulmón de adenocarcinoma basado en agrupamiento

Bryan Cervantes-Ramirez, Francisco Siles..... 16

Assessing the effectiveness of diarization algorithms in costa rican children-adult speech according to age group and gender

Evaluación de la efectividad de los algoritmos de registro en el habla de niños y adultos costarricenses según grupo de edad y género

Alejandro Chacón-Vargas, Daniel Pérez-Conejo, Marvin Coto-Jiménez..... 24

Exploring the potential of an audio application for teaching AI-based classification methods to a wider audience

Explorando el potencial de una aplicación de audio para enseñar métodos de clasificación basados en IA a una audiencia más amplia

Gabriel Coto-Fernández, Marvin Coto-Jiménez 33

Assessing the effectiveness of transfer learning strategies in BLSTM networks for speech denoising

Evaluación de la eficacia de las estrategias de aprendizaje por transferencia en las redes BLSTM para la reducción del ruido

Marvin Coto-Jiménez, Astryd González-Salazar, Michelle Gutiérrez-Muñoz 42

GPU based approach for fast generation of robot capability representations

Enfoque basado en GPU para la generación rápida de representaciones de capacidad de robot

Daniel García-Vaglio, Federico Ruiz-Ugalde 50

Personalized patient ventilation at large scale: Mass Ventilation System (MVS)	
Ventilación personalizada del paciente a gran escala: Mass Ventilation System (MVS)	
<i>Ogbolu Melvin Omone, Bence Takács, Roland Dóczy, Tivadar Garamvólyi, Lászlo Szücs, Péter Galambos, Tamás Haidegger, Miklós Vincze, Kristóf Papp, Daniel Drexler, György Eigner, Abdallah Benhamida, Ezer Koroknai, Peter Dombai, Miklos Kozlovsky.....</i>	58
Two-dimensional gel electrophoresis image analysis of two Pseudomonas aeruginosa clones	
Análisis de imágenes bidimensionales de electroforesis en gel de dos clones de Pseudomonas aeruginosa	
<i>José Arturo Molina-Mora, Diana Chinchilla-Montero, Carolina Castro-Peña, Fernando García</i>	67
Assessing costa rican children speech recognition by humans and machines	
Evaluación del reconocimiento de voz de los niños costarricenses por humanos y máquinas	
<i>Maribel Morales-Rodríguez, Marvin Coto-Jiménez</i>	74
Colorectal cancer vaccines: in silico identification of tumor-specific antigens associated with frequent HLA-I alleles in the costa rican Central Valley population	
Vacunas contra el cáncer colorrectal: identificación in silico de antígenos específicos de tumores asociados con alelos HLA-I frecuentes en la población del Valle Central de Costa Rica	
<i>Diego Morazán-Fernández, José Arturo Molina-Mora</i>	83
Automatic diagnosis of lower back pain using gait patterns	
Diagnóstico automático del dolor lumbar mediante patrones de marcha	
<i>Chandrasen Pandey, Neeraj Baghel, Malay Kishore-Dutta, Carlos M. Travieso.....</i>	93
Artificial Intelligence based Multi-sensor COVID-19 Screening Framework	
Inteligencia artificial para el marco de detección COVID-19 multisensorial	
<i>Rakesh Chandra-Joshi, Malay Kishore-Dutta, Carlos M. Travieso</i>	101
A deep learning approach for epilepsy seizure detection using EEG signals	
Un enfoque de aprendizaje profundo para la detección de ataques de epilepsia mediante señales de EEG	
<i>Manoj Kaushik, Divyanshu Singh, Malay Kishore-Dutta, Carlos M. Travieso</i>	110
Comparison of four classifiers for speech-music discrimination: a first case study for costa rican radio broadcasting	
Comparación de cuatro clasificadores para la discriminación de voz y música: un primer estudio de caso para la radiodifusión costarricense	
<i>Joseline Sánchez-Solís, Marvin Coto-Jiménez</i>	119
Application of Fischer semi discriminant analysis for speaker diarization in costa rican radio broadcasts	
Aplicación del análisis semi discriminante de Fischer para la diarización de locutores en transmisiones de radio costarricenses	
<i>Roberto Sánchez-Cárdenas, Marvin Coto-Jiménez.....</i>	128
A low cost collision avoidance system based on a ToF camera for SLAM approaches	
Un sistema de prevención de colisiones de bajo costo basado en una cámara ToF para enfoques SLAM	
<i>Dayron Romero-Godoy, David Sánchez-Rodríguez, Itziar Alonso-González, Francisco Delgado-Rajó.....</i>	137
A first study on age classification of costa rican speakers based on acoustic vowel analysis	
Un primer estudio sobre clasificación por edades de hablantes de costarricense basado en análisis de vocales acústicas	
<i>Victor Yeom-Song, Marvin Coto-Jiménez.....</i>	145
An experimental study on footsteps sound recognition as biometric under noisy conditions	
Un estudio experimental sobre el reconocimiento del sonido de las pisadas como biométrico en condiciones ruidosas	
<i>Marisol Zeledón-Córdoba, Carolina Paniagua-Peñaranda, Marvin Coto-Jiménez.....</i>	153

Presentation special issue

Presentación número especial

Melvin Ramírez-Bogantes¹, Jose Luis Vásquez-Vásquez²,
Carlos M. Travieso-González³

Ramírez-Bogantes, M; Vásquez-Vásquez, J.L;
Travieso-González, C.M. Presentation special issue. *Tecnología en Marcha*. Vol. 35, special issue. International Work Conference on Bioinspired Intelligence. November, 2022.
Pág 3-4.

 <https://doi.org/10.18845/tm.v35i8.6431>

Many specifics of the bioinspired intelligence are not well addressed by the conventional models currently used in the field of artificial intelligence. The purpose of the work conference is to present and discuss novel ideas, work and results related to alternative techniques for bioinspired approaches, which depart from mainstream procedures.

Nowadays, the studies based on complex system is opened new doors in research field and, to improve the quality and the results of diverse applications. The bioinspired intelligence is done easy this task and in areas like, biodiversity conservation, biomedicine, security applications, etc.

For this edition, the bioinspired intelligent have been applied in different areas, as the biomedicine, the speech and audio, the microbiology and the use of machine learning for different real applications.

In the area of biomedicine, six works will be presented. The first paper presents a combined approach with EP psychological assessments and EEG functional connectivity. The next one proposes an automatic concept-level neural network method to distilling genuine sentiment in patients' notes with Parkinson's Disease as medical polar facts into true positives and true negatives. The third document shows an approach toward lung cancer histological tissue images segmentation based on colour. The fourth work describes a Mass Ventilation System (MVS) which serves as a medical ventilator system. It can be used to ventilate large number of COVID-19 patients in parallel (5 – 50+) with personalized respiratory parameters. In other study, deep learning models can be trained with multiple chest x-ray images belonging to

1 Instituto Tecnológico de Costa Rica. Costa Rica. Email: meramirez@itcr.ac.cr

2 Universidad de Costa Rica. Costa Rica. Email:

3 Universidad de Las Palmas de Gran Canaria. Spain. Email: carlos.travieso@ulpgc.es

different categories to different health conditions i.e. healthy, COVID-19 positive, pneumonia, tuberculosis, etc. and finally, the last work proposes a one-dimensional Convolutional Neural network (CNN) for the automatic detection of epilepsy seizures.

Another area is the speech and audio, and six documents are included. The first one, authors perform an exploratory study with two diary algorithms in children-adult interactions within a recording studio and assess the effectiveness of the algorithms in different age groups and genders. The second paper, authors assess commercial include automatic speech recognition systems for the recognition of Costa Rican children's speech, for users with ages ranging between three and fourteen years old. The next work explores the application of several classifiers for the task of discriminating speech and music in Costa Rican radio broadcast. In the fourth work, a first annotated dataset and analysis of speaker diarization for Costa Rican radio broadcasting is performed, using two approaches: a classic one based on k-means clustering, and the more recent Fischer Semi Discriminant. The following paper presents the initial results on the identification of Costa Rican children's speech, in a database created for this purpose, consisting of words pronounced by adults and children of several ages. And finally, the last study is focused in the exploration of pure audio signals of footsteps and the robustness of a person's classification under noisy conditions.

The microbiology is studied by two works. In the first document, a classical strategy to analyse the protein content of a biological sample is the two-dimensional gel electrophoresis (2D-GE). This technique separates proteins by both isoelectric point and molecular weight, and images are taken for subsequent analyses. And finally, in the last study, authors propose a bioinformatic protocol to detect tumor-specific antigens associated with single nucleotide variants (SNVs) or "mutations" in colorectal cancer and their interaction with frequent HLA alleles (complex that present antigens to immune cells) in the Costa Rican Central Valley population.

Finally, the use bioinspired intelligence is applied in real applications by five studies. The first application, authors present a pilot project for teaching an AI-based classification method that is empirically evaluated with real data of a real problem. In the second document, a comparative study on different transfer learning strategies for reducing training time and increase the effectiveness of this kind of network is presented. The following study presents a method for generating capability maps taking advantage of the parallelization that modern GPUs offer such that these maps are generated approximately 50 times faster than previous implementations. The fourth work presents a machine learning method to diagnose these disorders using the Gait monitoring system. It involves support vector machines that classify between lower back pain and normal, on the bases of 3 Gait patterns that are integrated pressure, the direction of progression, and CISP-ML. And finally, the last paper designs an implementation and evaluation of a robust obstacle detection and mapping system.

As editors of this special issue, I would like to thank the authors; their effort and dedication that they have made to achieve some works of great quality. The sum of this effort has produced this special issue, which has become an inescapable read for all those who want to know the latest advances in bioinspired intelligence.

The impact of social media messages on Parkinson's disease treatment: detecting genuine sentiment in patient notes

El impacto de los mensajes de las redes sociales en el tratamiento de la enfermedad de Parkinson: detección de sentimientos genuinos en las notas de los pacientes

Hanane Grissette¹, El Habib Nfaoui²

Grissette H.; Nfaoui E.H. The impact of social media messages on Parkinson's disease treatment: detecting genuine sentiment in patient notes. *Tecnología en marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 5-15.

 <https://doi.org/10.18845/tm.v35i8.6441>

- 1 LISAC Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
E-mail: hanane.grissette@usmba.ac.ma
- 2 LISAC Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
E-mail: elhabib.nfaoui@usmba.ac.ma

Keywords

Distributed Biomedical representation; sentiment analysis; emotion modelling; patient narratives; social networks.

Abstract

Parkinson's Disease (PD), one of the most serious neurodegenerative diseases that known huge controversy on social networks. Following medical lexicons, few approaches have been extended to leverage sentiment information that obviously reflects the patient's health status in terms of related-narratives observations. It is been crucial to analyze online narratives and detect sentiment in patients' self-reports. In this paper, we propose an automatic concept-level neural network method to distilling genuine sentiment in patients' notes as medical polar facts into true positives and true negatives. Towards building emotional Parkinsonism assisted method from Parkinson's Disease daily narratives di- gests, we characterize polar facts of defined medical configuration space through distributed biomedical representation at the concept-level as- sociated with real-world entities, which are operated to quantifying the emotional status of the speaker context. We conduct comparisons with state-of-art neural networks algorithms and biomedical distributed systems. Finally, as a result, we achieve an 85.3% accuracy performance, and the approach shows a well-understanding of medical natural language concepts.

Palabras clave

Representación biomédica distribuida; análisis de sentimiento; modelado de emociones; Narrativas de pacientes; Redes sociales.

Resumen

Enfermedad de Parkinson (EP), una de las enfermedades neurodegenerativas más graves que ha suscitado una gran polémica en las redes sociales. Siguiendo los léxicos médicos, se han extendido pocos enfoques para aprovechar la información del sentimiento que obviamente refleja el estado de salud del paciente en términos de observaciones narrativas relacionadas. Es crucial analizar las narrativas en línea y detectar el sentimiento en los autoinformes de los pacientes. En este artículo, proponemos un método automático de red neuronal a nivel de concepto para destilar el sentimiento genuino en las notas de los pacientes como hechos médicos polares en verdaderos positivos y verdaderos negativos. Hacia la construcción del método asistido por el parkinsonismo emocional a partir de los diálogos narrativos diarios de la enfermedad de Parkinson, caracterizamos los hechos polares del espacio de configuración médica definida a través de la representación biomédica distribuida a nivel de concepto asociada con entidades del mundo real, que se operan para cuantificar el estado emocional del contexto del hablante. Realizamos comparaciones con algoritmos de redes neuronales de última generación y sistemas distribuidos biomédicos. Finalmente, como resultado, logramos un rendimiento de precisión del 85,3%, y el enfoque muestra una buena comprensión de los conceptos del lenguaje natural médico.

Introduction

Computational approaches are widely known surge application for alleviating Parkinson disease shared problems and experiences such as drug misuse or symptoms. Whereby, they used them for many cutting edge benefits such as mining shared experience to discover the risk of disease at an early stage whether a given patient has already developed or is expected to

develop Parkinson's disease in the future. Parkinson's Disease [1] Multiple examples of unsafe and incorrect treatment recommendations are shared every day. The challenge, however, to yield efficient, quick, consistent access to reliable insight. In the first sub-step, medical concept discrimination is proposed for data normalization [2], they automatically recognizing medical concepts mentioned in social media narratives enables several related-studies for enhancing health quality of people in a community, e.g. real-time monitoring of infectious diseases in the population, 1) we investigate a based-phrase transition between social media messages and formal description on medical anthologies such as MedLine.

Social Networks are frequently becoming popular as a platform for sharing personal health-related experiences [3] e.g. feedback, motivation, encouragement, emotional living, concrete support. Indeed, they support exchanged among those who go through emotional experiences too and usually, are influenced by the context surrounding the patient [4]. The advance of machine learning algorithms generally depends on data representation [5], and this information can be used for monitoring public health status. It hypothesized that these shared experiences might have distinct representations that can entangle and hide more or less the different descriptive factors of variation behind data and their properties. The advance of machine learning algorithms generally depends on data representation [5], neural networks based approaches such as Embeddings provide many advantages :(1) automatic feature extraction, (2) rich representation capabilities, (3) dimensionality reduction, (4) state-of-art performance for sentiment classification than tradition machine learning algorithms.

As many based neural networks approaches are becoming resolved many related medication text challenges, traditional based Lexicons methods fail requires an unaffordable huge labelled medical data or high qualified natural concept identification process. Due to such various factors as noised text errors and limited lexicons for related-medication text, the existing approaches cannot adequately make use of the sentiment information in the sentence for sentiment analysis. Many Distributions combined different pre-trained embeddings to enhance sentiment classification model performance, [6] proposed Sentiment Information Extractor based on Bi-directional Long Short Term Memory structure that applied to join the results of various sub-extractors. In particular scope, few approaches have been proposed for mining sentiment from related-medication text [7]. Patients notes on social networks and the use of informal medical language pose additional text challenges, including non-standard format, wrongly spelt, and abbreviation forms, as well as typos in social media messages. Nowadays, it is of crucial importance to providing an efficient solution to both mine medical components that extract aspects-based sentiment information. Most of the existing research, in this case, assumed at building a hand-annotated dictionary that is a time-consuming effort and may be subject to annotators/experts bias. However adopting an existing approach for sentiment analysis to mine related-medication texts may result in low classification accuracy, which employs less useful features sets and therefore lacks discriminatory potential. Many recent contributions aim at enhancing medical natural concepts processing by incorporating formal medical knowledge such as MedLine [8]. They developed a Neural network model, which further exploited N-gram-based convolutions for creating medical vocabulary scheme, which is dedicated mainly to featuring text under medical setting and clarifying related sentiment at the same level. In this paper, we provide an efficient method to both deploy medical concept-level conceptualization through generating real-world embeddings and define dynamic sentimental measure. Our main aims are as follow:

- Distill Parkinson's related-messages to treat what kind of PD's classes may uncover greater insights.
- Build an Emotion Parkinson's Assisted Tool to attempt a public concern about Parkinson's disease and their attributes.

- Propose an Inference component for evaluating note's positiveness/negativeness value, which measures the impact of given related-PD drugs or treatments.

The remainder of this paper is structured as follows. Section(2) provides a summary of the literature review concerning PD studies. Section(3) explains the offered methods. Section(4) offers large social network experimentation out- puts convolving large biomedical embedding. Section(5) presumes this paper and presents future aims.

Literature review

The accuracy of traditionally and potentially proposed clinical approaches [9] for PD classification, and prediction models have known slow progress due to data inaccessibility and complexity, where shows apparent diversity adopting these techniques on similar online narratives. They jointly processed varied techniques included data aggregation, rebalancing imbalanced cohorts, and meanings correspondence of heterogeneous data elements. They have been mostly integrated solutions from multiple sources offers unrivalled opportunities to track disease progress, e.g. early stages of prevalent neurodegenerative manners, and promptly discover the efficacies of alternative treatments based on the clinical definition of Parkinson's disease.

However, recent studies [10] hypothesize that it is likely to develop a computing model, which experiments results in acceptable accuracy, significantly, when he/she is anticipated to develop the Parkinson's disease to overcome data properties difficulties and challenges. Where artificial neural networks significantly outperformed the other methods in the literature and achieved the highest classification results reported so far. Most of the existing studies and PD methodologies rely on voice disorders patterns analysis; few works have been involved in leveraging self-reported experiences shared on social networks. Nevertheless, the results of several machine-learning based classification methods involved mandatory additional tasks, including Parkinson's disease evaluation metrics such as consistent accuracy, sensitivity, and statistical n-fold cross-validation. Indeed, it requires patterns recognition preliminary steps and medical preprocessing or normalization.

According to [11], the huge use and communications share of many social plat- forms such as AskPatient, DailyStrength, and Twitter, draw colossal attention from many researchers, which the potential for monitoring health conditions (e.g. adverse drug reactions, infectious diseases) in particular communities. They employed the power of the neural network to construct dynamic and credible Parkinson's Disease Digital Biomarker Dataset Using Neural Network Construction (NNC) Methodology Discriminates Patient Motor Status [12]. Nevertheless, these online shared experiences pose additional related-medical challenges. Refers to [8], daily patients' notes on social media may be a pivotal role to re- veal crucial public health issues through distilling hidden shared information, which remained so far unconsidered. Embeddings widely propose efficient dense representations that enhance sentimental analysis performance, which mostly focuses on discovering grammar directions such as semantic direction or only rely on upon represent sentiment property of distinct words at word-level. Sentence- level SA approaches to leverage the sentiment through the ensemble of tokens. In contrast, most of them have not capable of discovering sentiment conveyed towards related drug entities, or determine aspects-based sentiment within patients narratives. where a given sentence may contain complicated descriptive and nontechnical medical entities such as disease symptoms, drug misuse, and adverse drug effects.

However, The use of informal language while patients note concerning their treatments are characterized by a combination of nontechnical medical language with specific terminology (e.g. adverse effects, drug names) with greater lexical diversity. Deep learning methods have been used to cover these limitations. Most existing studies adopt word-mapping techniques to

normalize and translate related medical concepts and their attributes. [11] adopt a combination of a phrase-based Machine Translation(MT) method and the similarity between dense representations of words for overlapping social media phrase to the formal medical concept in standard ontologies; where it implicitly and explicitly succeed health quality monitoring in real-time window [2]. Moreover, Indeed, Medical sentiment analysis and emotion recognition have known by aspects' diversity. The previous study focused on different aspects, such as drugs and doctors' aspects [13]. At this end, Bi-directional Long Short Term Memory (BiLSTM) approach widely used as state-of-the-art methods to retrieve and classify sentiment within unstructured text [6] where the ensemble of convolutional approaches applied for medical recognition and extraction tasks.

Methods

The concept-level approaches are considered the key to common sense in neural network thinking. In this section, we seek to demonstrate the importance of establishing PD natural medical descriptive entities to the formal correspondence. Then, quantifying the emotional status of the speaker context. The way we can uncover the impact of social messages from shared Parkinson's Disease experiences through distilling genuine sentiment in patients' notes as polar medical facts into true positives and negatives narratives. In this section, we first present the rationale for this study case background. After, we address the proposed methods in detail.

Aspect-based distributed conceptualization

The concept-level approach is the key to provide an appropriate medical abstraction that is termed implicitly in all fundamental aspects of relationships among words, concepts, and medical definitions. Aspect-based conceptualization offers an integrative theory based on biomedical distributed representation. To the aim to find the relevant correspond of natural medical concepts and entities. Following existing Medical Sentiment analysis approaches, most of them fail due to the inability to extract related-medication concept aspects from the text.

Medical components normalization, from generated narratives on social networks, frequently needs medical lexical comprehension at the features extraction level, which can be difficult without sufficient background knowledge of the formal medical contexts. These explain why Sentiment Analysis systems can perform sentiment analysis towards a given entity and services reasonably well but far on clarifying sentiment towards a set of words that may refer to the specific medical components such as drug misuse multi-word expression. At this end, we choose to build a feature-based mechanism in terms of embedding vectors that incorporate external biomedical definitions and universal words for higher generalization performance. The embeds related-drug concepts preserve to define unrelated-medical items/ concepts to the formal medical language used in the descriptions of medical concepts in standard ontologies. The efficient use of pre-trained language medical knowledge guarantees a contextualized and semantical overlapping of primary medical observations in text. In the experiments, we used a data set that consists of two corpora for mining drug-related knowledge from online data, as illustrated in table 1.

Table 1. Statistics of distributed model for mining drug related knowledge from online data.

Distributed model	Description	Statistics
Drug-related, word representations from Twitter [14]	Distributed models based on distributional semantics capturing sequential patterns	It consists of 267,215 Twitter with over 250 drug-related posts mention keywords using drug names as keywords, including their common misspelled forms
Biomedical Word embeddings with subword information and MeSH (named BioWordVec) [15]	An open set of biomedical word vectors/ embeddings that combines subword information from unlabeled biomedical text with a widely-used biomedical controlled vocabulary called Medical Subject Headings (MeSH)	Biomedical Word and sentence embeddings using PubMed and the clinical notes from MIMIC-III Clinical Database as described in Table 2.

Table 2. Statistics of Controlled medical vocabulary we used.

Sources	Documents	Sentences	Tokens
Pubmed	28,714,373	181,634,210	4,354,171,148
MIMIC III Clinical notes	2,083,180	41,674,775	539,006,967
The PubMed Central (PMC)	2,083,180	41,674,775	539,006,967

Sentiment Language Modelling

Sentiment language modelling is the core of comprehending the genuine sentiment in patients' notes by quantifying the patient emotional status degree to state the social impact. It consists of an automatic online method that produces relevant aspects-based related-sentiment features, identifies and stores related- drug natural patterns from varied patient spaces, and then define the impactful class concerning medical aspects. So far, our sentimental analysis leverage differently concerning embed and contextual sentiment. It involved standard medical knowledge because it enables to encapsulate and parametrize new unobserved related-drug concepts throughout the downstream deep distributed medical components representations.

The relationship between emotion and language attributes to many challenges, especially, PD shared data has additional problems. Neural network methods have known massive distributions, where they mostly operated embeddings to get vectors of unknown raw data and define local semantics and contextual features to track related-sentiment value. Nevertheless, adopting the same manner on related-medication data results in slow accuracy than desired because of the model neglect medical entities' semantics. We propose to deploy both approximate modelling as follows :

- Generating biomedical Embedding vector built by adopting controlled vocabulary from medical ontologies and online version as described in Section 3.1.
- Re-calculate embeddings for unknown terms and medical relations.
- Computing statistical sentiment value regard medical aspects and targets defined Before.
- Constructing Our sentiment inference model based on Bidirectional LSTM, enhancing results by using an Attention mechanism.

The sentiment value is inferred regarding the relationships over words and concepts. we defined sentiment relationship by computing the pointwise mutual information (PMI) as quantified value that measure sentiment association within given words and/or concepts and the average concepts as follow:

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

By approximating the probabilities $p(x)$ and $p(x,y)$ respectively, we define sentiment association value with regard to biomedical embeddings. We present and discuss, in several different perspectives, the evaluation results of Machine learning models (LSTM, BiLSTM, SVM, simple RNN, linear regression) with based- associative biomedical embedding, and sentiment scale generation. The evaluation models have slightly different results, the most suitable algorithms that both leverage sentiment from semantics reasoning in varied contexts.

Implementation and Output Result

Keep track of health issues such as infectious diseases through Sentiment Analysis(SA) is of critical importance. Parkinson's disease(PD) long-term degenerative disorder that affects an estimated seven to 10 million people and families worldwide. Reports pour daily into healthcare communities, micro-blogs at a staggering rate, e.g. PD patients tend to be affected by severe emotional disorders, especially, PD's n-polar emotional state. In this paper, we ingest Parkinson's related messages from Michael J. Fox Foundation¹ and Twitter Data, to attain a better view of distinct drug or treatment in real-time.

Dataset and parameters

Twitter is the common popular micro-blogging that attracted millions of users to share and disseminate the most up-to-date information. In this study, we validated the proposed model by training our ensembles models using twitter data. We are also interested in evaluating the proposed method on extensive discussions from typical PD community with for evaluating how many methods comprehend varied particularities. Most of the PD communities are of public access, where PD patients and their families prefer to gather or share essential information about their minute treatments and problems from specific PD forums. In this paper, we ingest Parkinson's disease-related messages from the Michael J. Fox Foundation community to prove the efficiency of data on large forums that remained more details concerning drug and treatment vs disease. We collect data from this forum according to discussion topics and their attributes as described in table.4.

In our crawling system, data are collected using Twitter APIs. We used a list of related-PD Keywords Parkinson and dopamine. The table. 3 summarizes statistics of raw data in terms of the context of varied medical concepts and keywords. Parkinson communities have a particular structure; we choose categorized discussions into sub-forums, then into topics such as (1) Coronavirus (COVID-19) and Parkinson's for discussing the Corona virus pandemic, and how it coincides with Parkinson's disease. (2) Parkinson's Disease symptoms are uniquely challenging. For example, As PD symptoms are progressively damaged, their treatments description on social networks may be ambiguous and appear in the same manner in various discussions.

Table 3. Statistics of raw data in terms of the context of varied medical concepts and keywords.

Data set	Scheme and attributes	Keywords and statistics
Twitter Data	['id', 'created_at', 'source', 'original_text', 'retweet_count', 'original_author', 'hashtags', 'user_mentions', 'place', 'place_coord_boundaries']	15000 tweets collected regarding set of related-PD term and drugs e.g. <i>Tremor, bradykinesia, synuclein</i>
Parkinson's Disease(PD) forums	[Post_ids, topic, links, voices, replies, comments]	4807 original post by topics, subforum reference and their attributes.

Do, disambiguating related-entities for given related-PD discussions help to connect with others and access resources that can help explain, and cope with PD symptoms. As described, we streamed data regarding a set of keywords and topics. Then, raw data are stored to a set of _les regarding drugs vs disease components and time axis; we focused on associated criteria of chronic conditions, PD symptoms and related-medication concepts. Moreover, these samples were passing by a set of normalization and processing steps to be able for our neural inference model. Indeed, every single tweet in our dataset consists of a separate document that saves the life of correlate information contained regarding medical and pharma objects.

Table 4. Statistics of raw data in terms of the context of varied medical concepts and keywords.

Raw data topic	Statistics
PD & Dystonia	180 posts
CBD Oil and Parkinsons	115 posts
Butyric Acid and Niacin	97 posts
Has Covid-19 changed your doctor appointments?	98 posts
No pharmaceutical drugs?	121 posts
CRISPR GENE THERAPY! What do you know? Is it hope?	77 posts
Magnesium-L-threonate and PD	28 posts
How has your diet changed since the diagnosis?	45 posts

Implementation

A sample of 2807 PD-related posts, from the online healthcare community of Parkinson's disease, were collected and normalized to enrich the vocabulary. For twitter, we collect more than 15000 tweets that have been prepared to be input to the neural classifier for defining medical concepts and then re-define distributed representation for unrelated items of natural medical concepts cited in real-life patients narratives.

Experimentation and Results

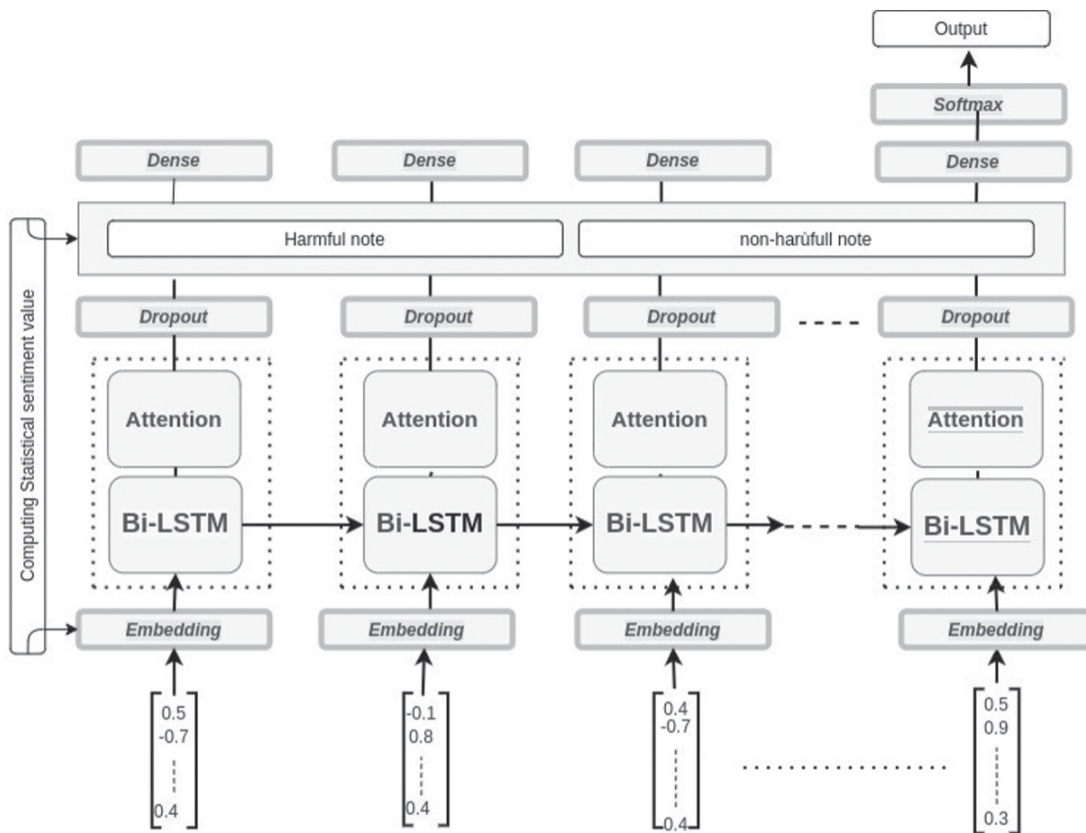


Figure 1. Sentimental inference model.

In this section, we present the results obtained by applying the proposed methods on both Twitter data and patients notes and messages on the forums. The experiments executed consider the language in which the messages are written, emoticons, mentions, and the length of texts analyzed.

The sentiment value is floated between $[-1, +1]$ -1 extremely negative +1 extremely positive. we create a novel sentiment scale for interpretable sentiment analysis model, which is dedicated to labelling positiveness and negativeness by harmful and non-harmful notes of given posts regarding related-medical targets, the way we can state false positives and negatives from texts. As described in figure 1. Each sentiment word belongs to a harmful class has to be convolved with an extremely negative value and clear medical co-annotation.

We conducted various experimentation originated from the application of the proposed method based on hybrid medical corpora based-text and concepts; a minute conceptualization is released. Sentiment detection is assessed through experiments on Three online datasets from various observations. An extensive evaluation of different features, including medical corpora, neural network algorithms. The based stacked-LSTM and BiLSTM model consistently improved the sentiment classification performance but is efficient when we exploit proposed configuration on a big online dataset as illustrated in table 5. We conduct development vocabulary scheme to enhance the effectiveness and transferability of this study across various online related-medication posts (Twitter posts, Parkinson's disease forum's posts), which was significantly better than all other baselines.

Table 5. Experiments results overview on Both Twitter Data and forums discussions (discussed in Section 3.1.)

Raw data	Neural Network algorithm	Accuracy
PD Tweets	LSTM	77%
	BiLSTM	85.3%
	Stacked-LSTM	79%
PD forums posts	LSTM	79.5%
	BiLSTM	87%
	Stacked-LSTM	78%

Discussion

Parkinson's disease (PD) is the second most crucial age-related disorder, with a prevalence ranging from 41 per 100,000 in the fourth decade of life to over 1900 per 100,000 in people over 80 years of age [16]. Refers to [12], the problem with Parkinson's disease (PD) patient monitoring fail is due to the limited resource. Where shared data suffered from inadequate, sporadic symptom monitoring, infrequent access, and sparsity, leading to poor medical decision making and sub-optimal patient health-related outcomes. Few Data Mining and NLP techniques have proposed in the PD context. As the existing tools are assisted for clinical decision-making, which is relying on the guidance of physicians to determine various measurement variables depends on diagnostic and treatments. In this study, we aim at conducting a demonstration of how can social media data can automatically leverage the social impact of PD patients and their families. Emotional dysregulation is an essential dimension that may occur in several psychiatric and neurologic disorders. We focused on clinical characteristics of emotional state variations in bipolar disorder and Parkinson's disease [17]. In both pathologies, the emotional intensity variability involves essential diagnostic and therapeutic issues for monitoring the emotional state of PD patients and track the impact of social media messages. Technically, instead of investing based-deep learning techniques to classify patients' self-reported messages on social media as positive or negative statements, we extend the sentiment inference model to distil sentiment aspects to distinguish negative facts degrees regarding Parkinson's disease-related drug-targets. Indeed, we are working on powered- Neural Network fact polar model to probe what kind of based-treatment target may result in improved Emotion Parkinson's model performance.

Likewise, harmful and non-harmful notes of given posts regarding related- medical targets noticeably may fail to retrieve the correct impact due to the inability to define complex medical components in text. Each post may refer to a drug reaction or/and misuse is categorized to harmful impact, where beneficial adverse reactions may also consider as harmful. As a solution, we propose to detect primarily drug reaction multi-expression in the text that should be considered as targets.

Conclusion

In this research, we studied a based neural network approach through multiple Bi-LSTM components to build a dynamic configuration space from unsupervised medical concepts representations. Thus, the embeddings from this joint model are used to generate sentiment scale powered by additional statistical salient for medical concept-aspects sentiment Inference.

In the future, we aim at defining related-PD drug related-events such as drug effectiveness, drug reaction, or drug misuse in text, which is assured to distinguish credible harmful insights.

References

- [1] Salama A. Mostafa, Aida Mustapha, Mazin Abed Mohammed, Raed Ibraheem Hamed, N. Arunkumar, Mohd Khanapi Abd Ghani, Mustafa Musa Jaber, and Shihab Hamad Khaleefah. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. *Cognitive Systems Research*, 2019.
- [2] Nut Limsopatham and Nigel Collier. Normalising medical concepts in social media texts by learning semantic representation. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016.
- [3] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 2015.
- [4] Rodrigues R.G., das Dores R.M., Camilo-Junior C.G., and Rosa T.C. SentiHealth- Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, 2016.
- [5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.
- [6] Kai Shuang, Zhixuan Zhang, Hao Guo, and Jonathan Loo. A sentiment information Collector–Extractor architecture based neural network for sentiment analysis. *Information Sciences*, 2018.
- [7] Hanane Grissette, EL Habib Nfaoui, and Adil Bahir. Sentiment Analysis Tool for Pharmaceutical Industry & Healthcare. *Transactions on Machine Learning and Artificial Intelligence*, 2017.
- [8] Hanane Grissette and El Habib Nfaoui. Enhancing convolution-based sentiment extractor via dubbed N-gram embedding-related drug vocabulary. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9(1):42, 2020.
- [9] Ivo D. Dinov, Ben Heavner, Ming Tang, Gustavo Glusman, Kyle Chard, Mike Darcy, Ravi Madduri, Judy Pa, Cathie Spino, Carl Kesselman, Ian Foster, Eric W. Deutsch, Nathan D. Price, John D. Van Horn, Joseph Ames, Kristi Clark, Leroy Hood, Benjamin M. Hampstead, William Dauer, and Arthur W. Toga. Predictive big data analytics: A study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS ONE*, 2016.
- [10] Niousha Karimi Dastjerd, Onur Can Sert, Tansel Ozyer, and Reda Alhajj. Fuzzy Classification Methods Based Diagnosis of Parkinson's disease from Speech Test Cases. *Current Aging Science*, 2019.
- [11] Nut Limsopatham and Nigel Collier. Adapting phrase-based machine translation to normalise medical terms in social media messages. In *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015.
- [12] Ioannis G. Tsoulos, Georgia Mitsi, Athanassios Stavrakoudis and Spyros Papapetropoulos. Application of machine learning in a Parkinson's disease digital biomarker dataset using Neural Network Construction (NNC) methodology discriminates patient motor status. *Frontiers in ICT*, 2019.
- [13] Salud María Jiménez-Zafra, M. Teresa Martín-Valdivia, M. Dolores Molina-González, and L. Alfonso Ureña-López. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. *Artificial Intelligence in Medicine*, 2019.
- [14] Abeed Sarker and Graciela Gonzalez. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. *Data in Brief*, 2017.
- [15] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWord-Vec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 2019.
- [16] Ramon Cacabelos. *Parkinson's disease: From pathogenesis to pharmacogenomics*, 2017.
- [17] Max Little, Paul Wicks, Timothy Vaughan, and Alex Pentland. Quantifying short-term dynamics of parkinson's disease using self-reported symptom data from an internet social network. *Journal of Medical Internet Research*, 2013.

Automated adenocarcinoma lung cancer tissue images segmentation based on clustering

Segmentación automatizada de imágenes de tejido de cáncer de pulmón de adenocarcinoma basado en agrupamiento

Bryan Cervantes-Ramirez¹, Francisco Siles²

Cervantes-Ramirez, B.; Siles, F. Automated adenocarcinoma lung cancer tissue images segmentation based on clustering. *Tecnología en marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 16-23.

 <https://doi.org/10.18845/tm.v35i8.6442>

- 1 Pattern Recognition and Intelligent Systems Laboratory (PRIS-Lab) and Cancer and Surgery Research Center (CICICA). Universidad de Costa Rica. Costa Rica. E-mail: bryan.cervantesramirez@ucr.ac.cr
- 2 Pattern Recognition and Intelligent Systems Laboratory (PRIS-Lab) and Cancer and Surgery Research Center (CICICA). Universidad de Costa Rica. Costa Rica. E-mail: francisco.siles@ucr.ac.cr
 <https://orcid.org/0000-0002-6704-0600>

Keywords

Digital pathology; pattern recognition; lung cancer.

Abstract

Cancer is one of the main dead causes worldwide. It is responsible for an approximate of 1 out of 6 deaths globally and lung cancer is along breast cancer, the most common types of cancer in the population, which confirms the importance of studies associated with it. This work presents an approach toward lung cancer histological tissue images segmentation based on colour. The proposed method for the segmentation is K-means clustering, providing promising results that may become as an assistance for pathologists, as it can help them reduce the time consumed reviewing the slides and giving a more objective perspective in order to provide a diagnose and specific treatment.

Palabras clave

Patología digital; reconocimiento de patrones; cáncer de pulmón.

Resumen

El cáncer es una de las principales causas de muerte en todo el mundo. Es responsable de aproximadamente 1 de cada 6 muertes a nivel mundial y el cáncer de pulmón, junto al cáncer de mama, es el tipo de cáncer más común en la población, lo que confirma la importancia de los estudios asociados a él. Este trabajo presenta un enfoque hacia la segmentación de imágenes de tejido histológico de cáncer de pulmón en función del color. El método propuesto para la segmentación es el agrupamiento de K-medias, brindando resultados prometedores que pueden convertirse en una ayuda para los patólogos, ya que puede ayudarlos a reducir el tiempo consumido revisando las diapositivas y dando una perspectiva más objetiva para brindar un diagnóstico y tratamiento específico.

Introduction

According to World Health Organization (WHO) data, cancer is the second cause of death worldwide and is responsible for an estimated 9.6 million deaths in 2018 [1]. Globally, about 1 in 6 deaths is due to cancer. In addition, for the particular case of lung cancer, it is estimated that it corresponds to a total of 2.09 million deaths of the aforementioned total, which makes it, together with breast cancer, the most common types of cancer in the population [1]. There are different types of tumors associated with lung cancer, such as non-small cell lung cancer (NSCLC) which is the case of the adenocarcinomas, small cell lung cancer (SCLC), carcinoid tumors, among others [2]. An accurate diagnosis is essential for lung cancer, since the spread of the disease and the response to treatment varies greatly between patients.

In general, when there is suspicion of the presence of some kind of cancer, the diagnosis is confirmed with a biopsy. A biopsy is a medical procedure that involves taking a small sample of tissue or cells to be examined in detail through a microscope [3]. Pathologists, based on the biopsy study, either diagnose cancer or, if it is already confirmed, evaluate its severity or current stage. They examine biopsy tissues through naked eye, looking for any visible abnormalities, and selecting certain regions of interest (ROIs) for further detailed analysis under the microscope. These small sections of tissue are stained with different types of chemicals that allow different

structures of the cells to be observed, in case they need to be evaluated in more detail. Histology is the study of the microscopic structure of tissues [4], and one of the most popular staining techniques used in histology is Haematoxylin and Eosin (H & E).

Microscopic examination of tissue slides is a crucial stage in the diagnosis of cancer and, simultaneously, the process is time-consuming, subjective and generates considerable interobserver and intraobserver variability. The interobserver variability occurs when different people evaluate the same case obtaining different results, while the intraobserver variability occurs when the same person evaluates a case in different moments in time, also differing in the results. Due to the subjectiveness and time-consuming nature of slide analysis by pathologists, the idea of automating the process at least to some extent sounds appealing, since having an image previously classified will benefit the pathologist, making it's labor faster and more objective. Of course, this tools would be produced as a complement to the work of the pathologist, and not as a substitution for it. Some approaches to automate the analysis of images from cell populations or histology slides, related to cancer cell tracking and classification, uses for example Hidden Markov Models [5], or classifies mitosis stages for phenotype classification [6]. Some other approaches more focused in the efficiency rather than the accuracy are based on convolutional neural networks (CNN) to achieve this classification, such as [7] and [8]. The approach showed in [9], which compares two CNN performance, reports an accuracy of 75%, and as it has been described previously, high accuracy is still crucial for diagnose. In addition, just for the adenocarcinoma cases, it is estimated that in 80% of cases there is presence of a wide mixture of histological patterns that must be qualitatively classified by a pathologist [10], which clearly shows the complexity of the task. Methods based on CNN be- have themselves as a 'black box', which creates a challenge in the extraction of relevant biological information and obtain meaningful insights from the trained models. There is still a need to clearly identify ROIs, in order to make a deeper analysis in specific areas of the issue and have an improvement in diagnose and treatment.

In the present work, the results of an algorithm for segmentation of bright- field microscopy lung cancer tissue images H&E-stained, based on clustering is presented. The clustering algorithm selected is K-means, which identifies different ROIs on the images associated to the dominant colors detected. In order to start working just with relevant areas of the image, a threshold was set in order to remove the background. Then, the histogram analysis will give a significant insight about the behavior of the color distribution in the image, which helps to get an idea about the potentially dominant colors.

Methods

In Figure 1, a simple diagram with the followed process is presented. Initially, the images are converted from SVS format to TIFF format, then is converted to HSV colour space in order to extract the ROI and finally apply K-Means algorithm.

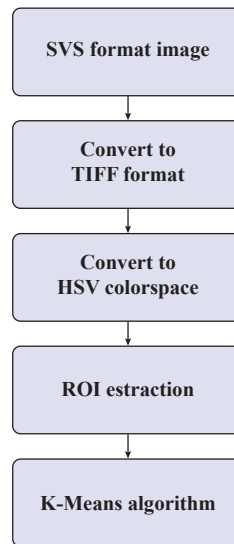


Figure 1. Overview of general process.

Dataset

All the images used for this work are adenocarcinoma images, which were obtained from the Clinical Proteomic Tumor Analysis Consortium Lung Adenocarcinoma (CPTAC-LUAD) group, which is part of the National Cancer Institute of the United States [11].

The images used for this work are ScanScope Virtual Slide (SVS) format. This format is proprietary from Aperio and it is specifically used for brightfield scanning. The slides are stored in a single file TIFF format, the first image in an SVS file is always the baseline image (full resolution). This image is always tiled, usually with a tile size of 240 x 240 pixels. The second image is always a thumbnail, typically with dimensions of about 1024x768 pixels. Unlike the other slide images, the thumbnail image is always stripped [12].

As these images are in a proprietary format, the first step is to convert them in TIFF format images so they can be processed with tools such as OpenCV [13]. In order to achieve this, the library ImageMagick was used.

It is important to mention that due to the conversion from SVS to TIFF format, in order to use OpenCV the maximum TIFF image size allowed is 1 GB. Nowadays scanned images in SVS format can easily go over 30-40 GB of storage, which brings a challenge in order to process any kind of algorithms on them.

ROI Segmentation

All of the scanned images have a white-tone background which is not relevant for the classification (see Figures 3a and 4a), so, in order to remove it, a threshold was defined as it is clearly showed that the background is not fully white. This threshold was selected experimentally, indicating in the mask, zeros in the pixels within the 'white' range.

The first step was to convert the original RGB image into HSV colour space, this was done because with classic RGB colour space it is more difficult to separate color information from elements such as lighting or intensity. So, after this conversion, the relevant channel is the hue one (H), as it has the color information. Then, using the threshold range previously explained, a mask was applied using an AND bitwise operation between the original HSV image and this mask.

Histogram analysis

The distribution of each of the pixels that make up the image was analyzed via histogram. As it was mentioned before, the relevant channel is the hue channel, as the desired result is to be able to discriminate each group of colors within the image, as one of each colors represent a different structure of the tissue that may be relevant for the pathologist analysis.

After visualizing the histogram for the hue channel, a series of considerations were made in order to select the steps to follow.

Based on the Central Limit Theorem, if each colour is considered as an independent random variable X_1, \dots, X_N , with overall mean μ , finite variance σ^2 and \bar{X} is the sample mean. Then the distribution of

$$Z = \frac{\bar{X}_N - \mu}{(\sigma/\sqrt{N})} \tag{1}$$

as $N \rightarrow \infty$, is the standard normal distribution [14].

After considering this, it is expected in the H channel histogram to get a mixture of gaussian distributions, made up of independent single distributions for each one of the dominant colors in the picture. In the histogram, this can be identified if multiple ‘peaks’ are being showed.

Segmentation

For the segmentation, a k-means clustering algorithm is used. The k-means is an unsupervised algorithm that it is used to assign to each pixel of the image one of the k-available tags with the degree of belonging to that particular cluster. One of the other parameters selected for the clustering was the amount of iterations, a total of 10 iterations was selected experimentally as a standard value for each test.

Results

The following results correspond to testing using K=3 and K=4 clusters. The amount of clusters selected was chosen based on naked-eye examination and identification on potentially amount of colors within each image. Also, as showed in Figure 2, the Elbow Method was used in order to get a first approach of which of the values may fit with the needs.

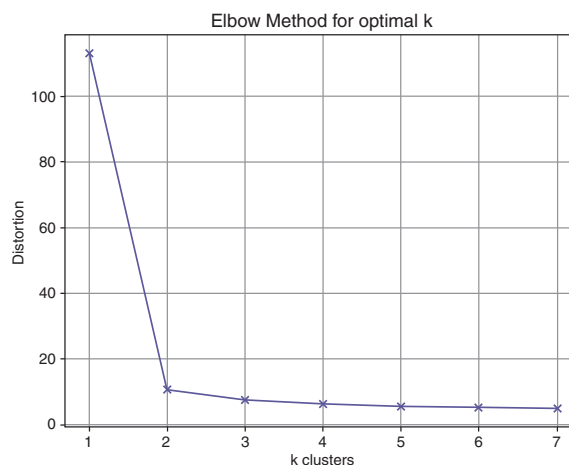


Figure 2. Elbow method to get optimal k.

Discussion and conclusions

The results show an appropriate segmentation within different structures of the tissue images. There is a clear challenge in the areas where the colour is very similar, and also whether to know if the number of clusters that are being suggested by the 'Elbow Method' is precise, since there might be no need to consider some of the areas of the slide that are being detected by the algorithm. The regions of interest for the pathologists may be just a small part of the complete tissue image that it is being processed, so as future work, after having previously annotated images by expert pathologists, then a complete validation of the algorithm can be performed, in order to have a more complete understanding on which parts are completely irrelevant and which ones are not.

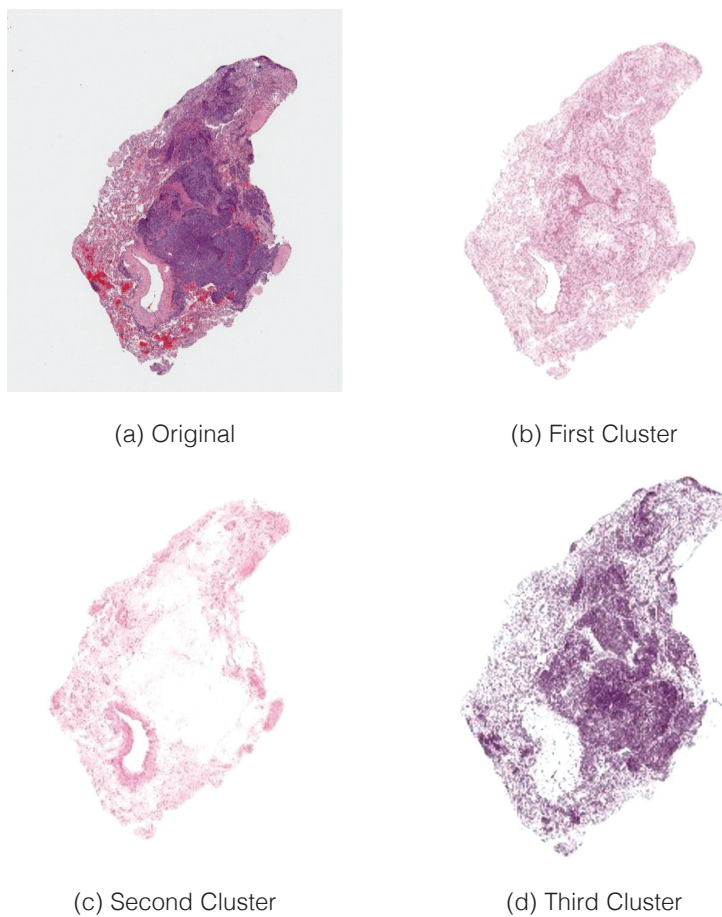


Figure 3. Original image and resulting clusters of the first slide.

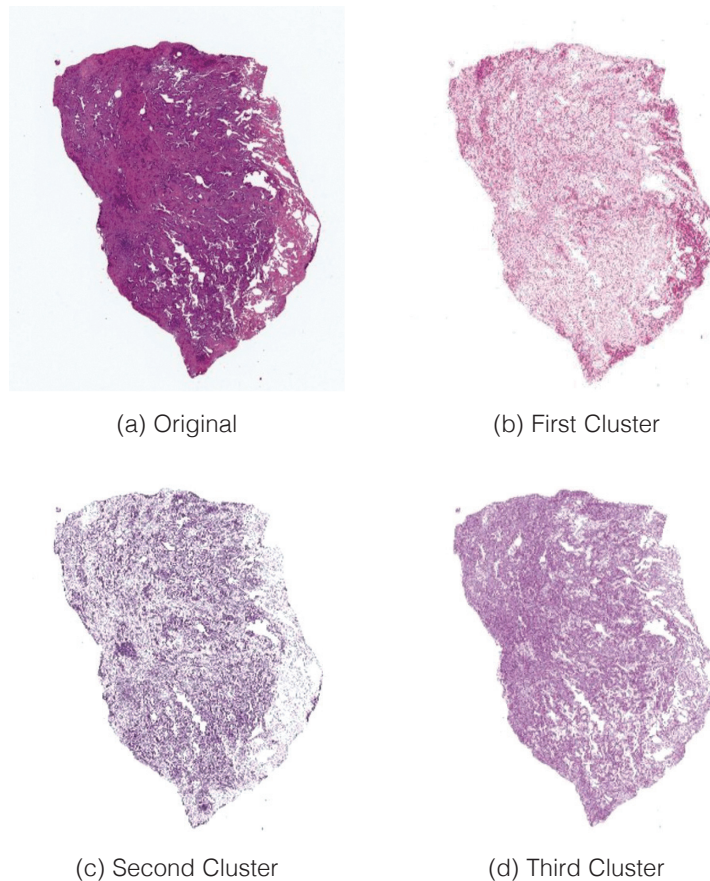


Figure 4. Original image and resulting clusters of the second slide.

In addition, this will create a better understanding of which other methods can be explored not just for the general area segmentation, but then into specific and deeper parts of it.

References

- [1] World Health Organization. Cancer key facts. <https://www.who.int/news-room/fact-sheets/detail/cancer>, Sep 2018.
- [2] American Cancer Society. What is lung cancer? <https://www.cancer.org/cancer/lung-cancer/about/what-is.html>, 2019.
- [3] The Royal College of Pathologists. What is a biopsy? <https://www.rcpath.org/discover-pathology/news/fact-sheets/what-is-a-biopsy.html>, 2018.
- [4] Brian Nation and Guy Orchard. Histopathology. Oxford University Press, 2 edition, 2012.
- [5] P. Quinde and F. Siles. Diseño, implementación y validación un algoritmo de rastreo de células cancerígenas basado en hmm's a partir de imágenes de microscopía de campo claro. Universidad de Costa Rica, 2018.
- [6] A. Mora and F. Siles. Cell phenotype classification using m-phase features in live-cell bright field time-lapse microscopy. Universidad de Costa Rica, 2018.
- [7] B. Peng, L. Chen, M. Shang, and J. Xu. Fully convolutional neural networks for tissue histopathology image classification and segmentation. In 2018 25th IEEE International Conference on Image Processing (ICIP), October 2018.
- [8] Y. Xu, Z. Jia, and L. Wang. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics, 2017.
- [9] M. Saric, M. Russo, M. Stella, and M. Sikora. CNN-based method for lung cancer detection in whole slide histopathology images. In 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), pages 1–4, 2019.

- [10] W.D. Travis, E. Brambilla, A.P. Burke, A. Marx, and A. G Nicholson. Who classification of tumours of the lung, pleura, thymus and heart, 2015.
- [11] National Cancer Institute. Clinical proteomic tumor analysis consortium lung adenocarcinoma. <https://wiki.cancerimagingarchive.net/display/Public/CPTAC-LUAD>, 2018.
- [12] OpenSlide. Aperio format. <https://openslide.org/formats/aperio/>.
- [13] G. Bradski. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 2000.
- [14] G. Cooper and C. McGillen. Probabilistic Methods of Signal and System Analysis. Oxford University Press, NJ, USA, 3rd edition, 1998.

Assessing the effectiveness of diarization algorithms in costa rican children-adult speech according to age group and gender

Evaluación de la efectividad de los algoritmos de registro en el habla de niños y adultos costarricenses según grupo de edad y género

Alejandro Chacón-Vargas¹, Daniel Pérez-Conejo², Marvin Coto-Jiménez³

Chacón-Vargas, A; Pérez-Conejo, D; Coto-Jiménez, M. Assessing the effectiveness of diarization algorithms in costa rican children-adult speech according to age group and gender. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 24-32.

 <https://doi.org/10.18845/tm.v35i8.6443>

1 University of Costa Rica. Costa Rica. E-mail: alejandro.chaconvargas@ucr.ac.cr

2 University of Costa Rica. Costa Rica. E-mail: daniel.perezconejo@ucr.ac.cr

3 University of Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr

 <https://orcid.org/0000-0002-6833-9938>

Keywords

Children's speech; clustering; speaker diarization; speech processing.

Abstract

Speaker diarization is the task of automatically identifying speaker identities and detecting their speaking times in an audio recording. Several algorithms have shown improvements in the performance of this task during the past years. However, it still has performance challenges in interaction scenarios, such as between a child and adult, where interruptions, fillers, laughs and other elements may affect the detection and clustering of the segments.

In this work, we perform an exploratory study with two diarization algorithms in children-adult interactions within a recording studio and assess the effectiveness of the algorithms in different age groups and genders. All participants are native Costa Rican Spanish speakers. The children have ages between 3 to 14 years, and the interaction combines guided repetition of words or short phrases, as well as natural speech.

The results demonstrate how the age affects the diarization performance, both in cluster purity and speaker purity, in a direct but non-linear fashion.

Palabras clave

Habla de los niños; agrupación; registro del hablante; procesamiento del habla.

Resumen

El registro de los oradores es la tarea de identificar automáticamente las identidades de los oradores y detectar sus tiempos de conversación en una grabación de audio. Varios algoritmos han mostrado mejoras en el desempeño de esta tarea durante los últimos años. Sin embargo, todavía presenta desafíos de desempeño en escenarios de interacción, como entre un niño y un adulto, donde las interrupciones, los rellenos, las risas y otros elementos pueden afectar la detección y agrupamiento de los segmentos.

En este trabajo, realizamos un estudio exploratorio con dos algoritmos de registro en interacciones niños-adultos dentro de un estudio de grabación y evaluamos la efectividad de los algoritmos en diferentes grupos de edad y géneros. Todos los participantes son hispanohablantes nativos de Costa Rica. Los niños tienen edades comprendidas entre los 3 y los 14 años, y la interacción combina la repetición guiada de palabras o frases cortas, así como el habla natural.

Los resultados demuestran cómo la edad afecta el rendimiento del registro, tanto en la pureza del grupo como en la pureza del hablante, de forma directa pero no lineal.

Introduction

Speaker diarization is often described as the task of automatically responding to the question of “who spoke when?” in an audio recording. The answer to the question, performed using a diversity of algorithms and approaches, is the time periods when each speaker is active, even for cases where the number of speakers is unknown.

Determining those time periods for each speaker can be useful in tasks, such as speaker indexing and the retrieval of large audio sets. Even transcription in movies, where automatically adding punctuation and speaker markers is of interest due to the large volume of films produced and

the possibility of making them accessible for people with disabilities [1]. Also, the segmentation produced with the time marks can benefit automatic speech recognition of more homogeneous segments, with a single speaker in each one.

The most typical scenarios are meetings, broadcast news and conversational telephone speech. In those cases, various problems need to be solved, including the building of clusters of information to estimate the number of speakers, and most of the times without a priori information, the processing of audios with music, silence, and other sounds, and the processing of spontaneous speech with overlapping voices of speakers [2].

More recently, the number of applications involving new scenarios and conditions have renewed interest in this task [3]; for example, the interaction between children and adults in clinical assessments and evaluations for children with autism spectrum disorder.

One of the reasons for the increasing interest of the research community in this topic is the availability of multimedia information in recent years. With this increase in the amount of information, more demanding conditions arose, for example multi-genre data, multiple speakers, diverse acoustic conditions or multiple recordings. A way to assess the progress and usefulness of this techniques is through evaluations in a particular domain [1].

Compared to the number of studies made for improving the performance of diarization systems, few studies have systematically analyzed the different sources and types of diarization errors [3].

In this work, we build an annotated corpus of children-adult interaction within a recording studio, and explore two diarization algorithms to determine systematically the performance and source of errors, especially those derived from the age range and gender of the children. To the best of our knowledge, this is the first study of speaker diarization for a population of Costa Rican children, which includes both guided repetition of words and phrases, and natural interactions.

As previous studies have reported in other languages, it is difficult for a child to focus his or her attention for long periods of time or remain in the same place. Thus, it is normal that the recording includes noises and interactions between the child and objects around [4]. For this reason, the diarization of children's speech is a particularly challenging task.

Related work

Among the many references on speaker diarization algorithms and conditions, in this section we describe those more closely related to the challenging conditions, and in particular the speech/adult interactions. For example, a recent experiment at the Johns Hopkins University [5] explored a variety of difficult conditions that demonstrate a high variability in the results, according to those conditions and the selection of features and procedures. Speaker overlap is an extra challenging condition for the algorithms [6], and is frequently found in recordings with children.

Most diarization methods take short segments of an audio recording and group the segments based on the clustering of a vector of parameters extracted from each one to conform to an i-vector (or fixed size vector). For example, such a procedure has been followed in [4] for the case of children's speech, using a Probabilistic Linear Discriminative Analysis with publicly available data for the English language. Some recent proposals have also considered additional elements of the recordings in order to improve the results, in English language [3]. Diarization of children and adult interaction for the English Language have been recently analyzed in [7]. In this case, the main concern is the analysis of the speech of children with an autism diagnosis. This kind of interaction is referred to as child-adult speaker classification in conversations with specified roles. In our work, a similar task is done for Costa Rican children, with the additional analysis of age and gender dependency.

Additionally, from collecting speech for speech technologies (e.g., speech synthesis and recognition) the procedure of diarization is of interest to language development researchers [8]. For example, determining the content of the child's language environment using a wearable recording unit, and categorizing the results according to whom initiated the speech and the number of interactions [9].

The challenges in the diarization of children's speech have also been reported in [10]. According to the authors, children's recordings may contain speech unfamiliar to usual processing models, such as cooing or crying. These kinds of elements can affect the algorithms that have proven efficiency with pure speech data. Additionally, the reference report that adults' speech when talking to infants has a larger pitch range.

Another factor that affects the accuracy of algorithms is the acoustic quality of the audio recording, distance between speaker and microphone and background noise. The accuracy of previous references using a single microphone located 1-2m from the speakers, achieved an accuracy rate of 70% [11].

In this work, we perform a first study on the diarization of children-adult guided interactions, using a state-of-the-art algorithm, with the purpose of verifying the conditions in the literature that affects the algorithms in this task, for the particular case of Costa Rican children. As we are focused on obtaining quality information from children of all ages, we classify the results according to that parameter and also the gender, to determine whether or not our interaction strategies are producing recordings suitable for automatic analysis.

The rest of this paper is organized as follows: Section 2 presents the experimental setup. Section 3 describes the results and discussion, and finally Section 4 presents the conclusions and future work.

Experimental Setup

In this section we summarize the main elements for building the database and the experiments performed to assess the diarization algorithms.

Database Recording and Processing

To obtain the recordings of the children-adult interactions, we conducted several recording sessions at a professional studio. The recordings were carried out using professional equipment, with one child and one adult each time. Those recordings were then manually edited and carefully annotated in terms of speaker turns and times, to define the ground truth for testing the algorithms.

Children Interaction

We established a pre-defined strategy with the use of pictograms designed to obtain repetitions of words with the children's voice. These recordings had a length of about 15 minutes for the children aged 8 to 14 years old, and about 10 minutes for children in the range of 3 to 7 years. Due to the limited time of concentration of some children, the interaction also considered some segments of jokes and laughs. Such elements are present more frequently in the recording of the younger children.

Diarization Algorithms

The Fisher Discriminant Analysis (FLD) is an algorithm for grouping vector of features. The criteria of the method is that the classes' means of each group are separated, and the variance within each group is small. According to the method describe in [12], a set of matrices is build, considering the mean of all vector of features m and the mean of all vector of features of each class m_c .

For example, the class scatter matrix is defined as

$$S_b = \sum_{c \in C} [(m_c - m) (m_c - m)^T] \quad (1)$$

where C is the set of all classes.

The average within-class scatter matrix S_v is defined as

$$S_v = \sum_{c \in C} \sum_{x \in C} [(x - m_c) (x - m_c)^T] \quad (2)$$

where x are the feature vectors.

The total scatter matrix of samples is defined as

$$S_m = \sum_{\text{all } x} [(x - m) (x - m)^T] \quad (3)$$

The aim of FLD is to find a matrix A that optimize a criterion of grouping separation in an optimal subspace.

The case of Fisher Semi Discriminant Analysis is applied when the set of classes are unknown. The idea, presented by [13] is to assume that neighbor samples of an audio file are likely to be part of the same class.

Evaluation

According to previous references [13, 14], the common measures for the effective- ness of a speaker diarization are the cluster purity and speaker purity, defined as follows:

- **Cluster purity:** The cluster purity is the percentage of data in each cluster which belongs to the most dominant speaker, according to the ground-truth (annotated) reference. Mathematically, it can be expressed as:

$$\text{Cluster purity} = \frac{1}{N} \sum_{i=1}^{N_c} \max_{j=1, \dots, N_s} n_{ij}, \quad (4)$$

where N is the total number of detected segments, N_s is the number of speakers, N_c is the total number of clusters, and n_{ij} is the total number of segments classified in cluster i and spoken by speaker j .

- **Speaker purity:** The speaker purity is the percentage of data of the most common detected speaker within each speaker class. This is a measure for the assessment of the speaker turns, and can be expressed as:

$$\text{Speaker purity} = \frac{1}{N} \sum_{i=1}^{N_s} \max_{j=1, \dots, N_c} n_{ij}, \quad (5)$$

where each symbol follows the previous description.

Results and Discussion

This section introduces the results of the diarization algorithms for the children speaking sessions in the developed database. The results are organized comparing FLSD algorithm with the simplest method of K-means.

In regards to cluster purity, the results in relation to participants' age are shown in table [1] for K-means algorithm. It is noted that it tends to the increase of the cluster purity as children age progresses. In some cases, they will increase or decrease, but they always tend to increase. Equally, Speaker Purity tendency tends to increase with ages, where it increases or decreases.

A similar trend is noted in Table [2] for the FLSD algorithm. In the case of Cluster Purity, a total of seven age ranges presents better results than the K-means case. As to Speaker Purity, the comparison with k-means case is for improvement: there are a total of eight cases that give a better result. Among the reasons that can be pointed out for the decrease of K-means and FLSD algorithms' capacity for carrying out diarization is that there are more elements of noise, short interactions, interruptions, and other speak fillers in younger children's interaction.

A comparison of the results according to children's gender is observed in graphics [1] and [2]. A tendency to obtain better results in boys than in girls is observed. This can be partly explained by the fact that the adult person who interacts with the children in all sessions is a woman; thus, it is more likely that there is a higher contrast with the boys' voices that allow a high contrast for the algorithm.

Table 1. Diarization results for K-means.

Session	Age	Number of speaking turns	Cluster purity	Speaker purity
1	3	120	84.4	51.7
2	4	125	84.4	51.7
3	6	117	83.3	73.7
4	6	103	74.0	74.0
5	7	101	72.7	72.7
6	7	114	72.9	57.6
7	8	96	79.3	79.6
8	11	144	72.4	59.1
9	11	68	80.2	73.2
10	12	170	72.3	50.0
11	12	53	83.6	83.6
12	14	49	98.2	97.4

Table 2. Diarization results for FLSD.

Session	Age	Number of speaking turns	Cluster purity	Speaker purity
1	3	120	84.4	41.6
2	4	125	84.4	41.6
3	6	117	74.3	74.3
4	6	103	83.3	82.5
5	7	101	79.2	79.2
6	7	114	72.9	70.8
7	8	96	79.3	79.0
8	11	144	77.8	77.8
9	11	68	80.2	73.7
10	12	170	72.8	52.1
11	12	53	85.1	85.1
12	14	49	98.5	95.0

As is observed in these results, there is a clear dependency of diarization algorithm capacity, according to the participants' age and gender. The use of these algorithms for this application environment does not seem to have dependable levels, so higher explorations must be realized, especially for the cases of younger female children.

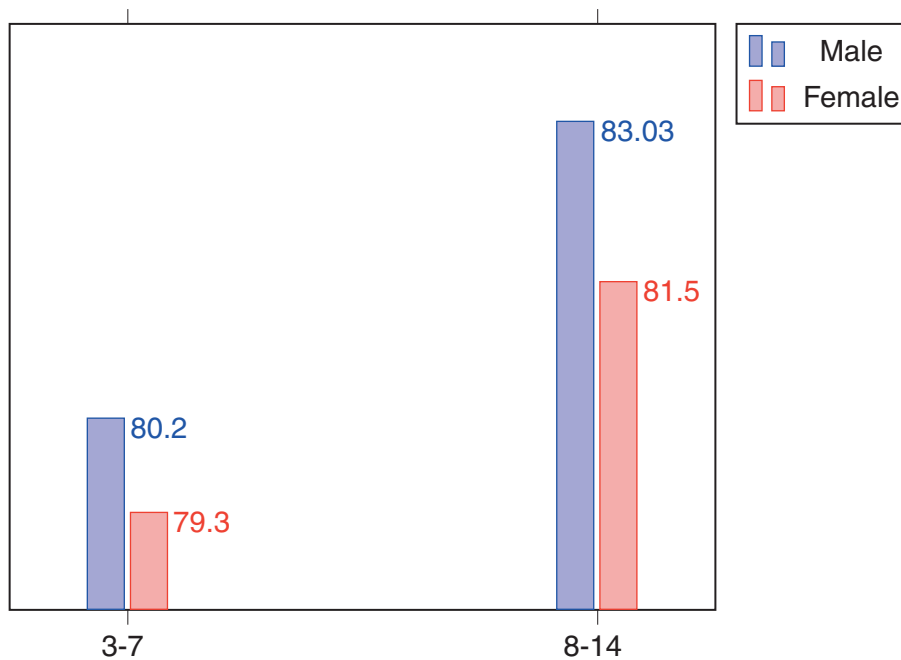


Figure 1. Cluster purity by gender (FLSD algorithm)

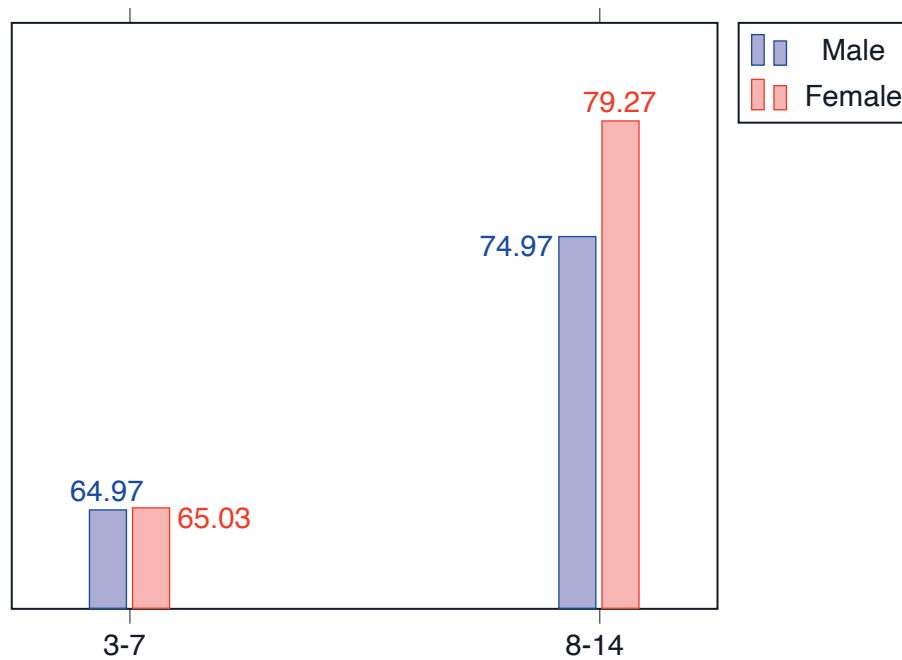


Figure 2. Speaker purity by gender (FLSD algorithm)

Conclusions

The purpose of this research is to determine the effectiveness of two diarization algorithms in relation to the age and gender of children. We perform the analysis in children-adult interactions with native Costa Rican Spanish speakers.

The results show that there is a direct correlation between the effectiveness of the algorithm and the age and gender of children. According to age, both K-means and FLSD algorithms tended to perform better in older children, and in relation to gender, male children obtained better results.

The results presented here, open up the possibility of exploring new specific strategies of diarization according to the age and gender of children, in order to improve the segmentation and grouping processes of speech in this type of sessions.

In regards to future work, it is intended to debug existing algorithms to improve their performance with young children, new accents, and noisy environments. It can also be explored with a greater combination of algorithms to obtain higher levels of reliability.

References

- [1] Karanasou, Penny, et al. "Speaker diarization and longitudinal linking in multi- genre broadcast data." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.
- [2] Meignier, Sylvain, et al. "Step-by-step and integrated approaches in broadcast news speaker diarization." *Computer Speech & Language* 20.2-3 (2006): 303-330.
- [3] Kumar, Manoj, et al. "Improving speaker diarization for naturalistic child-adult conversational interactions using contextual information." *The Journal of the Acoustical Society of America* 147.2 (2020): EL196-EL200.
- [4] Xie, Jiamin, et al. "Multi-PLDA Diarization on Children's Speech." *Interspeech*. 2019.
- [5] Sell, Gregory, et al. "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge." *Interspeech*. 2018.

- [6] Fujita, YusukeRao, et al. "Meta-Learning for Robust Child-Adult Classification from Speech." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
- [7] Koluguri, Nithin Rao, et al. "Meta-Learning for Robust Child-Adult Classification from Speech." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
- [8] Najafian, Maryam, and John HL Hansen. "Speaker independent diarization for child language environment analysis using deep neural networks." 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016.
- [9] Zhou, Tianyan, et al. "Speaker diarization system for autism children's real-life audio data." 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2016.
- [10] Karadayi, Julien, Camila Scaff, and Alejandrina Cristià. "Diarization in Maximally Ecological Recordings: Data from Tsimane Children." SLTU. 2018.
- [11] Gorodetski, Alex, Ilan Dinstein, and Yaniv Zigel. "Speaker diarization during noisy clinical diagnoses of autism." 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019.
- [12] Sarafianos, Nikolaos, Theodoros Giannakopoulos, and Sergios Petridis. "Audio-visual speaker diarization using fisher linear semi-discriminant analysis." *Multimedia Tools and Applications* 75.1 (2016): 115-130.
- [13] Giannakopoulos, Theodoros, and Sergios Petridis. "Fisher linear semi-discriminant analysis for speaker diarization." *IEEE transactions on audio, speech, and language processing* 20.7 (2012): 1913-1922.
- [14] Chen, Liping, et al. "On Early-stop Clustering for Speaker Diarization." *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*. 2020.

Exploring the potential of an audio application for teaching AI-based classification methods to a wider audience

Explorando el potencial de una aplicación de audio para enseñar métodos de clasificación basados en IA a una audiencia más amplia


Gabriel Coto-Fernández¹, Marvin Coto-Jiménez²

Coto-Fernández, G.; Coto-Jiménez, M. Exploring the potential of an audio application for teaching ai-based classification methods to a wider audience. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 33-41.

 <https://doi.org/10.18845/tm.v35i8.6444>

¹ Saint Joseph High School. Costa Rica

² PRIS-Lab, Electrical Engineering Department. University of Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr

 <https://orcid.org/0000-0002-6833-9938>

Keywords

Audio; artificial intelligence; classification; teaching.

Abstract

Knowledge about artificial intelligence (AI) is becoming increasingly important for many careers, especially those based in science and engineering. Besides formal education, the impact of AI on society lead to consider educational projects for teaching the fundamental concepts of AI at wider audiences, including high school levels. This can help more general audiences to better understand how AI works, with the hope that also parents and educators can help students develop a healthy appreciation for implications and limitations, along with an appropriate relationship and deeper interest on it. In this paper, we present a pilot project for teaching an AI-based classification method that is empirically evaluated with real data of a real problem, which can be understood and tackled with basic mathematical tools and activities suitable for high school students. With this proposal, we aim to show how audio and speech applications can inform a wider audience about advances in AI, its characteristics, and its future impact on society. Results and lessons learned from this project can form the basis for further projects using different tools and data, according to students' interests and initiative.

Palabras clave

Audio; inteligencia artificial; clasificación; docencia.

Resumen

El conocimiento sobre inteligencia artificial (IA) se está volviendo cada vez más importante para muchas carreras, especialmente las basadas en la ciencia y la ingeniería. Además de la educación formal, el impacto de la IA en la sociedad lleva a considerar proyectos educativos para enseñar los conceptos fundamentales de la IA a un público más amplio, incluidos los niveles de secundaria. Esto puede ayudar a un público más general a comprender mejor cómo funciona la IA, con la esperanza de que también los padres y educadores puedan ayudar a los estudiantes a desarrollar una apreciación saludable de las implicaciones y limitaciones, junto con una relación adecuada y un interés más profundo en ella. En este artículo presentamos un proyecto piloto para la enseñanza de un método de clasificación basado en IA que se evalúa empíricamente con datos reales de un problema real, que se puede entender y abordar con herramientas y actividades matemáticas básicas adecuadas para estudiantes de secundaria. Con esta propuesta, nuestro objetivo es mostrar cómo las aplicaciones de audio y voz pueden informar a una audiencia más amplia sobre los avances en IA, sus características y su impacto futuro en la sociedad. Los resultados y las lecciones aprendidas de este proyecto pueden formar la base para otros proyectos utilizando diferentes herramientas y datos, de acuerdo con los intereses y la iniciativa de los estudiantes.

Introduction

Artificial intelligence (AI) has become part of our everyday life: its existence, recent applications and potential impact are constantly discussed in mass media. Children now interact with tablets, toys and video games that implement AI algorithms, all of which contribute to their understanding of these devices. In this way, their knowledge about AI becomes more deeper and more nuanced [1]. Despite the importance of AI, few people know about the technology and its requirements [2]. This is one reason it is important to familiarize young people in their school years with the technical details and the underlying concepts of AI.

The concept of AI literacy has been well discussed in popular media outlets. It is expected that literacy in AI will become a major issue in the future. With

AI literacy, students can receive a solid preparation for subsequent studies at the university level and in their future careers.

According to reported studies, children can be too trusting of technological devices and can be influenced by them [3]. Better understanding of the principles and implications of AI in high schools and by parents and educators can help them teach children to gain an appreciation for AI's abilities and limitations, as well as build a proper relationship with it.

Some of the limitations that exist for the inclusion of teaching AI concepts in high schools and other wider audiences have been due to the complexity of the mathematical concept of algorithms, the computing and programming background needed to understand AI concepts, the many difficulties in the implementation of projects with a proper understanding of the implications and the environment where real applications should be developed.

Given those difficulties, it becomes very important to carefully select situations and teaching strategies that motivate new audiences and provide them an adequate level of understanding and appreciation for AI. Based on previous experience, some themes have been shown as key to teaching AI to children and wider audiences [4, 5]: 1) AI systems are based on knowledge acquired from human beings through examples. 2) AI systems do not know everything and make mistakes. 3) AI systems are corrected and improved by human beings.

Considering these key themes and the importance of real-world applications that can be understood and addressed by younger learners and general audiences, we propose the exploration of audio and speech in teaching strategies on AI concepts, especially those based on classification methods. For this purpose, we selected a real-world problem that arose from the need to help people with disabilities, and the opportunity to provide them with alternative and augmentative communication systems [6].

This problem led us to design an educational experience to explore the possibilities for teaching basic classification methods, including the complete process of building a data-set, presenting and explaining feature extraction, and visually applying the classification concepts.

Related Work

Several researchers have reported successful strategies for teaching AI concepts to children and other general audiences. For example, in [7], an approach for an agriculturally-based AI challenge that helped students learn the process of creating machine learning models was presented. The study found that machine learning can be used as a tool to conduct interdisciplinary education at the middle school level.

With a historical perspective, the authors in [8] discuss the Turing Test as an educational activity for undergraduate students. Some other long-term courses also introduce basic concepts, vocabulary and history [9] at appropriate levels, including topics presented in important books commonly used for graduate students, but adapted for younger students [2].

Associations such as the Association for the Advancement of Artificial Intelligence (AAAI) and the Computer Science Teachers Association (CSTA) in United States of America have formed working groups to develop national guidelines for teaching AI, machine learning and robotics to K-12 students. One of the main objectives of the projects is to invite the AI research community to reflect their ideas and developments of AI in a way that every K-12 student should know, and how communicate with the general public about advances in AI and their future impact on society [10].

In Latin America, the idea of presenting AI concepts to a wider audience have been discussed in [4], with the particular aim of an exhibition in a science museum.

All this initiatives apply different strategies to particular age groups, and focus on concepts of AI that suit some of the vast possibilities of education. Our approach take advantage of these experiences in term of do's and don'ts, situations to avoid and to consider, but with a new application based on audio and speech processing for helping people with disabilities.

The rest of this paper is organized in the following way: Section II presents the motivation that lead the proposal, in terms of developing communication solutions to people with disabilities. Section III shows the experimental results of the experience, Section IV the results and finally Section V present the conclusions.

Experimental and teaching framework

Motivation

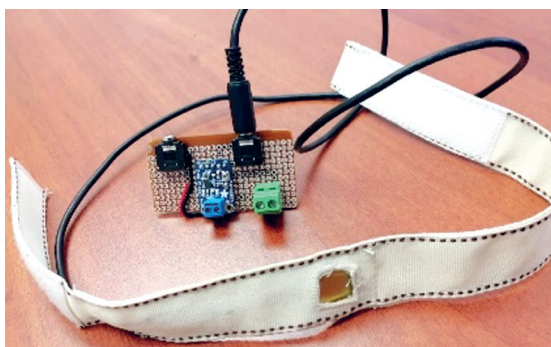
The motivation for the application of speech in this proposal arose from the fact that speech is the most common means of communication between people. However, some disability conditions may limit the abilities of people of different ages, genders and others conditions to speak. Within this framework, an AI-based application of speech technologies can provide opportunities for communication, cognitive support and functional independence for many users.

For example, some persons with brain paralysis cannot express themselves verbally, but are capable of producing sounds that express emotional states or positive/negative answers to questions. A simple device that incorporates a piezo-electric sensor and sound amplifier can be built with the appropriate characteristics for its implementation by taking audible vocalizations directly from the user's throat, and recording them with a computer or embedded system.

The AI-based classification algorithm can label new sounds according to the labeled sounds in a database. In the final stage, speech synthesis can pronounce the word in a a manner that can be understood by most people.

This motivation can be presented to high school students that can also participate in the making of the simple sensor, as shown in Figure 1.

Voluntary students or participants themselves can participate in the recording of the database, consisting on isolated gesticulations to express "yes" or "no", as described in the next section.



(a) Piezo-electric sensor and amplifier



(b) Recording sessions

Figure 1. Illustration of the proposed device.

Database description

Our proposal is to present the topic of AI-based classification methods with the necklace and the piezoelectric sensor describe in II.A, for the purpose of producing an alternative communication device for people with disabilities. To test this device and its ability to classify sounds with positive or negative meaning, the proposal is to record at least two volunteers producing three hundred and sixty sounds, with the following characteristics:

1. Thirty sounds indicating affirmation and thirty denial, with the head facing forward.
2. Thirty sounds indicating affirmation and thirty denial, with the head turned as far as possible to the right.
3. Thirty sounds indicating affirmation and thirty denial, with the head turned as far as possible to the left.

According to our experience, the recording sessions for this database can last about 20 minutes per volunteer, which represents a proper time for an experience suited for a wider audience. The head turns are proposed to verify the robustness of the device against involuntary movements, or the movements of continuous use in a realistic environment. During process, the feedback of the students or users can be of great benefit and can lead to discuss and record new data or to systematize the sessions.

For this work, we perform a recording session with volunteers, as a mean to properly measure the difficulties of the process and to validate the proposal. Feature extraction and results are reported to set expected values.

Feature extraction

The process of production of speech or sounds with the sounding apparatus is complex, since it involves a series of organs, from the lungs, the larynx, the vocal cords, the nose and the tongue, among others [11]. This complexity is illustrative in explaining the students the complexity of the speech production process and the subtle differences that let human beings differentiate voices and words.

This complexity also allows the introduction of how numerically characterize the sounds of the database. Whilst the parametrization of speech production is a complex subject for high school or more general audiences, the understanding of pitch, related to deep and sharp voice, or the melody produce in talking, can be shared and applied in this experience.

To maintain the necessary simplicity in the process, the classification of the sounds in the database is proposed using basic statistics of a single parameter: the fundamental frequency. This decision is based on the fact that there are numerous implementations for the extraction of the parameter, in addition to the simplicity of the process to properly implement a recognition system for the application of assistive technologies. There are also open research lines to detect it in adverse conditions [12], which is useful for real applications.

For each of the sounds in the database, the fundamental frequency and four characteristics are obtained:

- Min (minimum value of f_0 of the while file)
- Max (maximum value of f_0 of the while file)
- Mean f_0 value of f_0 of the while file)
- Median f_0 value of f_0 of the while file)

For evaluation purposes, we propose a subset of about 20 recordings (with the corresponding four statistical characteristics) to test the algorithm applied.

Algorithm and exploration

For a proper visualization and understanding of the AI-based classification scheme, we intend to apply the parameters in pairs, for a 2D representation of each file in the plane. One of the first questions that arose in real-life experimentation with classification problems is: What features should we use? This question can motivate discussion and experimentation with students.

For example, we can choose some of the pairs of values for a manual application of an algorithm such as KNN, where only distances between points need to be measured or estimated from a planar representation. In particular, the case of 1-NN can be presented and explained with ease at the high school level, where the notions plane, points and coordinates are well known.

The following steps can be applied in an teaching experience, after the introduction of the problem and an explanation of the 1-NN algorithm:

1. Plot the pairs of values of each possible combination of features in a plane, with corresponding labels for the negative or positive message of the audio.
2. Ask the students or participants to manually label each of the points of the test set according to its nearest neighbour.
3. Keep a second plot of each pair of features with the real labels of the points in the test set, for comparison purposes.
4. Compute the errors manually and decide on the best pair of features.
5. Compare the results with a computer version of the algorithm, where further experimentation can be done; for example, 2-NN, 3-NN and 4-NN.

The necessity of AI in this problem can be presented in terms of the difficulties that arise when elements that are always different (such as words and human sounds) need to be classified according to preset categories.

Results and discussion

This proposal can be useful in a teaching experience if the results represent a meaningful challenge for the participants and the error rates are close to those of a real application. For example, if the error rate is as high as 50%, the whole experience cannot be considered successful, nor will it lead to appreciation of the classification method and AI methodologies.

To verify the possibilities of the data-set to achieve good results, as well as manual labeling of the plots, in Figures 2 and 3 two pairs of features are shown, with the corresponding training and testing data, with spaces for manual labeling presented with lines.

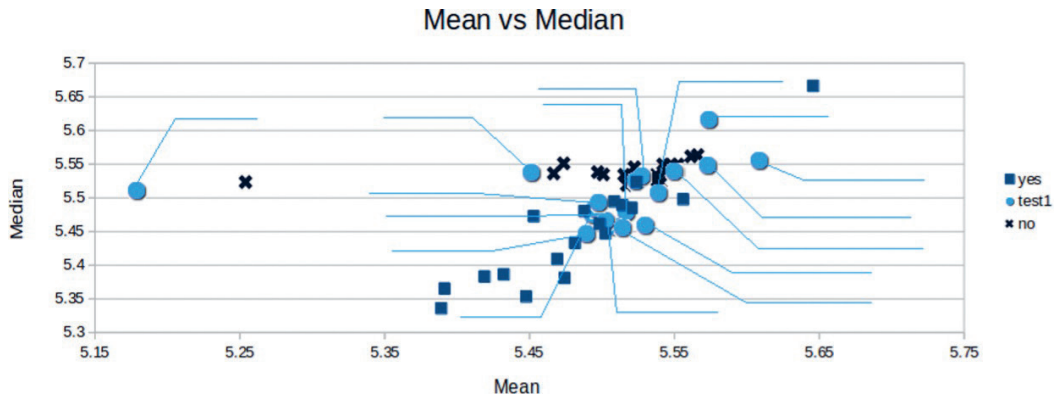


Figure 2. Set of points of the train and test data for Mean and Median features.

The two plots also show how different the classification method can be in terms of how points are distributed and separated in the training set. Some of the points of the test set are very close in both cases, and proximity to its nearest neighbor can foster collaborative work and discussion from participants.

Table 1 shows the results of a preliminary manual labeling and comparison with the KNN algorithm implemented in the Weka system, which can be very useful for demonstrations in teaching, due to its flexibility and user-friendly interface.

Many questions can be asked about the results shown in the plots and the table; for example: can human labeling surpass computer results? How much time does it take to prepare the files (such as the ARFF format for the Weka system) in comparison with manual labeling of results? What other features can be extracted from the dataset (such as range or quartiles)? Is it easy to find the best parameters for a classification method? Can computers make mistakes in classification or other AI-related tasks? How can these errors be minimized?

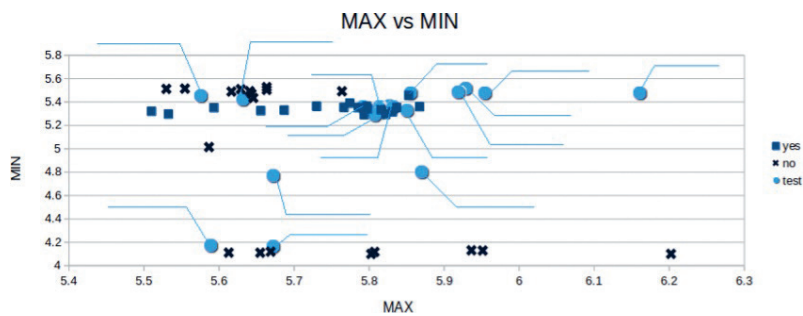


Figure 3. of points of the train and test data for Max and Min features.

Table 2. t -test results comparing ex-combatants and civilians on SCB scores.

Features	1-NN (Human)	1-NN	2-NN	3-NN	4-NN
Max vs Min	20	25	30	30	30
Mean vs Min	10	10	15	10	10
Mean vs Max	40	40	30	30	40
Mean vs Median	10	20	25	10	15

Further exploration with the algorithms, beyond what is shown in Table 1, can be easily done with the software. For example, the application of more than two features at the same time, and even the results from other, unknown algorithms that can surpass the obtained results, can show the wide variety and the importance of more exploration and study of AI.

Conclusions

In the present work, we explored the application of a real-life problem, related to a communication device for people with disabilities, in order to present a teaching experience of AI-based classification methods to a wider audience, who has some basic statistic knowledge and understanding of a few mathematical concepts, such as those studied in the first years of high school. For this purpose, we show a low-cost device that can be understood or even built for the participants, and the methodology to present and contextualize the problem, as well as the necessity of AI for the classification of sound recordings through the device.

Almost all of the problems of a real-life application of AI can be explored with the proposal, including recording of the data-set, parametrization, feature extraction and selection of the algorithm, all adapted and simplified for a wider audience. The results of a preliminary experience show the impact of feature selection on the results, and the meaningful experience that a manual application of the algorithm and the comparison to the results and the computer represent. For future work, we aim to develop high quality materials (such as large scale plots and presentations) to implement the proposal with a group of high school students. There is a wide margin for experimentation, from the materials and characteristics of the device, as well as the development of the database and extended conditions of use that can also be explored.

References

- [1] Williams, Randi, et al. "Popbots: Designing an artificial intelligence curriculum for early childhood education." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.
- [2] Burgsteiner, Harald, Martin Kandhofer, and Gerald Steinbauer. "Irobot: Teaching the basics of artificial intelligence in high schools." Thirtieth AAAI Conference on Artificial Intelligence. 2016.
- [3] Williams, Randi, et al. "My doll says it's ok" a study of children's conformity to a talking doll." Proceedings of the 17th ACM Conference on Interaction Design and Children. 2018.
- [4] Candello, Heloisa, et al. "30 Minutes to Introduce AI to Kids." (2019).
- [5] Kandhofer, Martin, et al. "Artificial intelligence and computer science in education: From kindergarten to university." 2016 IEEE Frontiers in Education Conference (FIE). IEEE, 2016.
- [6] González-Salazar, Astryd, Michelle Gutiérrez-Muñoz, and Marvin Coto-Jiménez. "Enhancing Speech Recorded from a Wearable Sensor Using a Collection of Autoencoders." Latin American High Performance Computing Conference. Springer, Cham, 2019.
- [7] Sakulkueakulsuk, Bawornsak, et al. "Kids making AI: Integrating Machine Learning, Gamification, and Social Context in STEM Education." 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE). IEEE, 2018.

- [8] Torrey, Lisa, et al. "The Turing Test in the classroom." Thirtieth AAAI Conference on Artificial Intelligence. 2016.
- [9] Heinze, Clint Andrew, Janet Haase, and Helen Higgins. "An action research report from a multi-year approach to teaching artificial intelligence at the k-6 level." First AAAI Symposium on Educational Advances in Artificial Intelligence. 2010.
- [10] Touretzky, David, et al. "Envisioning AI for K-12: What Should Every Child Know about AI?." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.
- [11] Holmes, Wendy. Speech synthesis and recognition. CRC press, 2001.
- [12] Coto-Jiménez, Marvin, John Goddard-Close, and Fabiola Martínez-Licona. "Improving automatic speech recognition containing additive noise using deep denoising autoencoders of LSTM networks." International Conference on Speech and Computer. Springer, Cham, 2016.

Assessing the effectiveness of transfer learning strategies in BLSTM networks for speech denoising


Evaluación de la eficacia de las estrategias de aprendizaje por transferencia en las redes BLSTM para la reducción del ruido

Marvin Coto-Jiménez¹, Astryd González-Salazar², Michelle Gutiérrez-Muñoz³


Coto-Jimenez, M., González-Salazar, A., Gutiérrez-Muñoz, M. Assessing the effectiveness of transfer learning strategies in BLSTM networks for speech denoising. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence Pág. 42-49.

 <https://doi.org/10.18845/tm.v35i8.6448>


1 Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: marvin.coto@ucr.ac.cr

 <https://orcid.org/0000-0002-6833-9938>

2 Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: astryd.gonzalez@ucr.ac.cr

 <https://orcid.org/0000-0002-3444-0464>

3 Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: michelle.gutierrezmunoz@ucr.ac.cr

 <https://orcid.org/0000-0003-3313-8324>

Keywords

BLSTM; deep learning; speech processing; transfer learning.

Abstract

Denosing speech signals represent a challenging task due to the increasing number of applications and technologies currently implemented in communication and portable devices. In those applications, challenging environmental conditions such as background noise, reverberation, and other sound artifacts can affect the quality of the signals. As a result, it also impacts the systems for speech recognition, speaker identification, and sound source localization, among many others. For denosing the speech signals degraded with the many kinds and possibly different levels of noise, several algorithms have been proposed during the past decades, with recent proposals based on deep learning presented as state-of-the-art, in particular those based on Long Short-Term Memory Networks (LSTM and Bidirectional-LSMT). In this work, a comparative study on different transfer learning strategies for reducing training time and increase the effectiveness of this kind of network is presented. The reduction in training time is one of the most critical challenges due to the high computational cost of training LSTM and BLSTM. Those strategies arose from the different options to initialize the networks, using clean or noisy information of several types. Results show the convenience of transferring information from a single case of denosing network to the rest, with a significant reduction in training time and denosing capabilities of the BLSTM networks.

Palabras clave

BLSTM; aprendizaje profundo; procesamiento del habla; aprendizaje por transferencia.

Resumen

La eliminación de ruido de las señales de voz representa una tarea desafiante debido al creciente número de aplicaciones y tecnologías implementadas actualmente en los dispositivos portátiles y de comunicación. En esas aplicaciones, las condiciones ambientales desafiantes como el ruido de fondo, la reverberación y otros artefactos de sonido pueden afectar la calidad de las señales. Como resultado, también afecta a los sistemas de reconocimiento de voz, identificación de hablantes y localización de fuentes de sonido, entre muchos otros. Para eliminar el ruido de las señales de voz degradadas con los muchos tipos y posiblemente diferentes niveles de ruido, se han propuesto varios algoritmos durante las últimas décadas, con propuestas recientes basadas en el aprendizaje profundo presentadas como vanguardistas, en particular las basadas en redes de memoria a corto plazo (LSTM y LSMT bidireccional). En este trabajo se presenta un estudio comparativo de diferentes estrategias de transferencia de aprendizaje para reducir el tiempo de formación y aumentar la efectividad de este tipo de redes. La reducción del tiempo de entrenamiento es uno de los desafíos más críticos debido al alto costo computacional de entrenar LSTM y BLSTM. Esas estrategias surgieron de las diferentes opciones para inicializar las redes, utilizando información limpia o ruidosa de varios tipos. Los resultados muestran la conveniencia de transferir información de un solo caso de eliminación de ruido de la red al resto, con una reducción significativa en el tiempo de entrenamiento y las capacidades de eliminación de ruido de las redes BLSTM.

Introduction

In the speech signal processing, clean audio is ideally expected, without additive or convolutional noise. Though, in uncontrolled conditions, audio signals are degraded by multiple unknown agents or conditions that affect the quality of speech. As a result, these conditions do not allow the optimal performance of speech technologies [1–3].

To solve this problem and enhance the signal quality recorded in real-life environments, several algorithms have been developed throughout the years. In this work, a method based on deep neural networks (DNN) was proposed to map the noisy speech to clean speech [4, 5]. As mentioned, with degraded signals, speech technologies do not work properly, for this reason, the DNN approach can be implemented for better results in several applications, such as in mobile phone applications, speech recognition systems, and assistive technology [6, 7].

One example of this successful method of noise reduction is the Long Short-term Memory (LSTM) neural networks and its bidirectional extension (BLSTM), which are models of recurrent neural networks (RNNs) [8]. Particularly in speech recognition, LSTM has shown better results than DNN or convolutional networks [9, 10]. On the other hand, their training procedures represent a high computational cost. For this reason, a study presented in [8] explained the advantages of using mixed neural networks for reducing computational cost in the task of reverberant speech enhancement. In this work, a comparative study is presented on different transfer learning strategies to improve the capacity of BLSTM neural networks for noise reduction and reducing training time in a set of different noise types and levels.

Transfer learning is a concept used in artificial neural networks (ANN), to improve the results of a model in one domain by transferring information from a model in a related domain. This can be described using a given source domain DS with a corresponding task TS, and a target domain DT with a corresponding task TT. The process improves the target predictive function $f_T(\cdot)$ by using the related information from DS and TS [11].

One of its advantages is that it can be applied with a reduced amount of training data in DT [12]. For this work, the homogeneous transfer process was applied, because there is available data that is drawn from a domain (denoising speech degraded with artificial noise) related to, but not an exact match for a target domain of interest (denoising speech degraded with natural noise).

Problem Statement

The modeling of speech signals degraded with background noise can be presented in its simplest way as the combination of a pure speech signal, x , with an additive noise d . Thus, the noisy signal $y(t)$ can be expressed in the time domain by the sum:

$$y(t)=x(t)+d(t) \quad (1)$$

In discrete frequency-domain, the relation is simply

$$Y_k(n)=X_k(n)+D_k(n) \quad (2)$$

For the statistical implication of several signal processing-based methods, the clean signal $x(t)$ may be modeled as independent and uncorrelated to the noise $d(t)$. This way, the algorithms can attempt to extract the information of $X_k(n)$ from $Y_k(n)$ and $D_k(n)$, and then reconstruct the approximation for $x(t)$.

From the perspective of deep learning approaches that apply artificial neural networks, $x(t)$ (or $X_k(n)$) can be estimated in the form of an approximated function $f(\cdot)$ that is calculated directly from samples of noisy and clean data of the form:

$$\hat{x}(t) = f(y(t)) \quad (3)$$

How well the task is performed by the artificial neural network, such as BLSTM, depends on the amount of available data and the type of network selected [13]. In our work $f(\cdot)$ is obtained not only from the pairs of data $y(t)$ and $x(t)$ but from another function estimated to perform a similar task, with other kinds of noises. This way, transfer learning is performed between two artificial neural networks.

To verify the benefits of this approach, several objective measures were applied, to compare the training of the BLSTM in the task of denoising several types of noise at several SNR levels.

Experimental setup

The experimental setup to test the Transfer Learning strategies in BLSTM Networks can be condensed in three steps:

1. Speech dataset generation: Two kinds of noise, one artificially generated (White Noise), and one naturalistic (the crowd noise Babble) were added to each clean utterance recorded in the dataset for different signal-to-noise ratio (SNR -10 dB, SNR -5 dB, SNR 0 dB, SNR 5 dB and SNR 10 dB), with the aim of covering light to heavy noise distortion in White and Babble noise.
2. Feature extraction: Owing to previous experiences in denoising autoencoders, MFCC information was extracted from each frame of the clean utterances and those degraded with noise, using the Ahocoder system [14]. Thus, pairs of noisy and clean parameters were presented to the networks as inputs and outputs during training.
3. Training and testing: The BLSTM networks are trained using backpropagation through time algorithm, to adjust the internal set of values in the connections according to the pairs of values presented to each BLSTM network. The total database of about 900 utterances was splitted 80% for training, 15% for validation and 5% for the test (about 50 files). Details and equations of the training algorithm followed can be found in [15].

Dataset

SLT (female) was selected from the CMU Arctic database [16], widely applied in speech research. This dataset were designed and produced at the Carnegie Mellon University. The SLT set of utterances consists of 1132 sentences, and split the whole set randomly to establish the training, validation, and test set, as described in the previous section.

Transfer Learning Strategies

One base system (training without transferring information) and three Transfer Learning strategies were compared:

- *Case 1 (Base system)*: This case corresponds to the traditional random initialization of the internal weights of the BLSTM network. The procedure were applied to all cases in order to establish the base system for comparison of the proposals.
- *Case 2 (Transfer from AAN)*: In this case, an auto-associative network (AAN) is trained from clean speech parameters. AAN approximates the identity function by presenting the same information at both the inputs and outputs. The set of internal weights resulting from this training are transferred to all the BLSTM network to begin its adjustment.

- *Case 3 (Transfer from White Noise)*: In this approach, a BLSTM network is trained for denoising the signal with SNR0 White Noise, and then the results are transferred to the whole set of BLSTM networks, both for denoising White and Babble Noise at all SNR levels.
- *Case 4 (Transfer from Babble Noise)*: Similar to the previous case, but the first BLSTM autoencoder is trained for denoising SNR0 of Babble Noise, and then the results are transferred to the whole set of BLSTM networks, both for denoising White and Babble Noise at all SNR levels.

The experimental study aims to numerically compare and recommend the best procedure in terms of less training time, whilst the capacity of the network in denoising the speech increases or is not affected considerable, according to the objective quality evaluation measures.

Evaluation

The following common measures for artificial neural network training were applied to the results given by the four cases of initialization and transfer learning:

- *SSE (sum of squared errors)*: Is the sum of the squared differences between each output and the expected value in the validation set, calculated at each training epoch. For a given network θ , SSE is computed as:

$$SSE(\theta) = \sum_{n=1}^T (c_x - \hat{c}_x)^2, \quad (4)$$

where c_x is the real value of the parameters in the validation set ($x(t)$ in Equation (1)), \hat{c}_x is the predicted output from each frame of the validation set ($\hat{x}(t)$ in Equation (3)), and T the total number of frames from each audio file.

- *Number of epochs*: An epoch is a complete cycle of the parameters in the dataset presented to network in order to adjust the update the internal weights of the BLSTM networks. The total time taken to train any neural network is the sum of the time taken by individual epochs required in the process.

All the BLSTM networks were trained on a Linux computer with Pentium i7 Processor accelerated with an Nvidia GPU.

Results

In table 1, the results of each of the four cases are presented. The missing value in the Transfer-White results is because this particular SNR level was used to initialize the rest of the BLSTM autoencoders, so there no information transferred from any other network to this SNR level.

Table 1. SSE results and number of epochs required for training (in parenthesis) for each case of White Noise analysed. * is the best result for each SNR level.

Case 1			Case 2	Case 3	Case 4
SNR-10					
Random 1 464.45(271)	Random 2 472.70(159)	Random 3 458.45(205)	Transfer-AAM 463.06(196)	Transfer-White* 447.94(96)	Transfer-Babble 457.52(211)
SNR-5					
Random 1 401.41(168)	Random 2 396.32(180)	Random 3 395.52(188)	Transfer-AAM 392.07(207)	Transfer-White* 380.63(162)	Transfer-Babble 392.17(168)
SNR0					
Random 1 332.67(196)	Random 2 328.65(269)	Random 3 332.64(262)	Transfer-AAM* 328.28(374)	Transfer-White -	Transfer-Babble 330.99(194)
SNR5					
Random 1 239.98(198)	Random 2 283.25(362)	Random 3 290.52(221)	Transfer-AAM* 283.22(394)	Transfer-White 285.40(157)	Transfer-Babble 288.82(275)
SNR10					
Random 1 251.86(376)	Random 2 251.80(340)	Random 3 253.68(333)	Transfer-AAM* 243.79(655)	Transfer-White 245.95(417)	Transfer-Babble 253.61(295)

According to the results presented in Table 1, in SNR-10 level, transferring the set internal weights from white noise and babble noise showed better results than from auto-associative memory and random initialization. Additionally, they obtained the result in less time and this is reflected in the number of epochs (case 1). Similarly, in SNR-5 the best result was obtained using Transfer-White: the smallest SSE value and training time (case 3).

As to the noise levels from SNR0 and on, the best results were obtained using AAM. However, once again the results obtained with transfer (from both white and babble noise) outperform the results obtained with random initialization. Additionally, in the transfer cases, the total training time of the networks is reduced, giving these an advantage in computational cost.

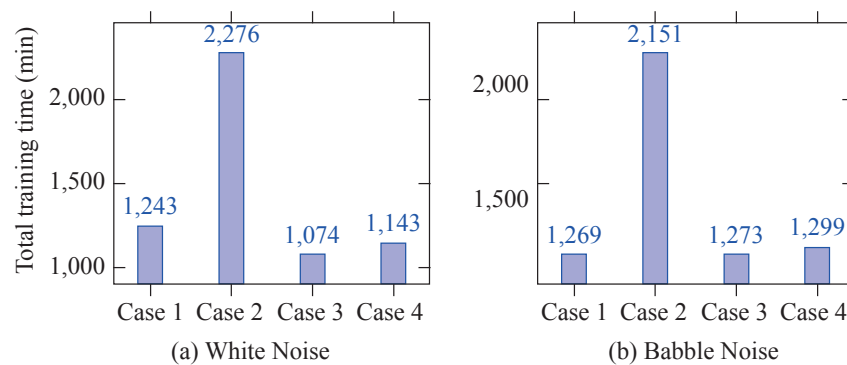


Figure 1. Total training time for denoising BLSTM networks

Transfer using AAM requires an additional training process that is not used for the denoising task but just for initialization purposes. Due to this additional training process, the total training time grows. In general, the good results obtained by the AAM came as a trade-off with training time.

The Figure 1a shows that the biggest average training time of the networks is case 2: 2151 min. This training time is almost 83.11% more in comparison to random initialization (case 1). In contrast, in cases 3 and 4, the results were better than case 1 and are competitive with case 2 in some levels, requiring a notably short training time between 47.19%-50.22% less than case 2.

In Table 2, the results of each of the four cases are presented. The missing value in the Transfer-Babble results is because this particular SNR level was used to initialize the rest of the BLSTM networks, and the best result was obtained transferring from auto-associative memory for four of five levels of noise; however, the computational cost is higher when compared with other cases. This is shown graphically in Figure 1b.

An advantage of using non-random procedures is that it is no longer necessary to use multiple trials to obtain significant good results compared to others, because of the great variability present in random initializations. For example, in Table II, for SNR-10 there are ranges from 585.85 to 603.75 for SSE values in Case

1. Instead, with the transfer procedures, the results obtained are competitive and (in some cases even better) with lower training time.

Conclusions

In this work, an experimental approach was performed to compare the efficiency of training denoising BLSTM networks, for natural and artificial noises. The results allow the numerical validation of the transfer learning procedure for regression problems in this kind of network, to establish that the random and independent initialization of a set of BLSTM networks is not the best option in terms of efficiency and training time. Transferring the weights from one BLSTM trained with a particular condition (such an SNR level) to initialize the weights of the rest of the conditions and noise types can reduce training time and increase efficiency.

Table 2. SSE results and number of epochs required for training (in parenthesis) for each case of Babble Noise analysed. * is the best result for each SNR level.

Case 1			Case 2	Case 3	Case 4
SNR-10					
Random 1 603.74(127)	Random 2 595.22(129)	Random 3* 585.85(139)	Transfer-AAM 586.36(163)	Transfer-White 593.13(102)	Transfer-Babble 595.17(67)
SNR-5					
Random 1 491.71(122)	Random 2 488.73(165)	Random 3 494.40(129)	Transfer-AAM* 466.43(185)	Transfer-White 499.29(79)	Transfer-Babble 481.43(47)
SNR0					
Random 1 374.24(175)	Random 2 370.51(164)	Random 3 372.14(189)	Transfer-AAM* 356.54(234)	Transfer-White 369.20(220)	Transfer-Babble -
SNR5					
Random 1 289.60(245)	Random 2 282.32(279)	Random 3 285.17(302)	Transfer-AAM* 266.12(587)	Transfer-White 283.73(317)	Transfer-Babble 280.34(400)
SNR10					
Random 1 221.24(636)	Random 2 230.20(484)	Random 3 225.41(522)	Transfer-AAM* 219.36(590)	Transfer-White 220.42(587)	Transfer-Babble 221.62(609)

In particular, the transfer of information from the BLSTM trained with white noise present the best results. It can be explained for the most homogeneous distribution of values that represent the white noise, but further experimentation should be conducted in order to validate this hypothesis. The transfer of information from Auto-associative Memories presents the best results in most cases, but the time required for its training is considerably higher.

It can be stated that random initialization for the complete set of neural networks is in no case the best option. Transfer from a particular network is the best option in terms of time and results, which can be the best option in exploratory studies of regression, such as exploring architectures or comparing neural networks. The transfer from Auto-associative Memories can be considered only to achieve the best results when the training time is not an issue.

For future work, more extensive validation of the transfer learning among some types of noise or particular SNR levels can be performed. Statistical validation of the improvements achieved could be relevant, along with numerical validation of the results in terms of the quality of the signal.

References

- [1] Weninger, F., Watanabe, S., Tachioka, Y., and Schuller, B. "Deep recurrent denoising auto-encoder and blind de-reverberation for reverberated speech recognition." IEEE ICASSP, 2014.
- [2] Donahue, Chris, Bo Li, and Rohit Prabhavalkar. "Exploring speech enhancement with generative adversarial networks for robust speech recognition." IEEE ICASSP, 2018.
- [3] Coto-Jiménez, Marvin, John Goddard-Close, and Fabiola Martínez-Licona. "Improving automatic speech recognition containing additive noise using deep denoising autoencoders of LSTM networks." International Conference on Speech and Computer. Springer, Cham, 2016.
- [4] Abouzid, Houda, et al. "Signal speech reconstruction and noise removal using convolutional denoising audio-encoders with neural deep learning." Analog Integrated Circuits and Signal Processing 100.3 (2019): 501-512.
- [5] Ling, Zhang. "An Acoustic Model for English Speech Recognition Based on Deep Learning." 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, 2019.
- [6] Coto-Jiménez, M.; Goddard-Close, J.; Di Persia, L.; Rufiner, H.L. "Hybrid Speech Enhancement with Wiener filters and Deep LSTM Denoising Autoencoders." In Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOB), San Carlos, CA, USA, 18–20 July 2018; pp. 1–8.
- [7] González-Salazar, Astryd, Michelle Gutiérrez-Muñoz, and Marvin Coto-Jiménez. "Enhancing Speech Recorded from a Wearable Sensor Using a Collection of Autoencoders." Latin American High Performance Computing Conference. Springer, Cham, 2019.
- [8] Gutiérrez-Muñoz, Michelle, Astryd González-Salazar, and Marvin Coto-Jiménez. "Evaluation of Mixed Deep Neural Networks for Reverberant Speech Enhancement." Biomimetics 5.1 (2020): 1
- [9] Tkachenko, Maxim, et al. "Speech Enhancement for Speaker Recognition Using Deep Recurrent Neural Networks." International Conference on Speech and Computer. Springer, Cham, 2017.
- [10] Liu, Ming, et al. "Speech Enhancement Method Based On LSTM Neural Network for Speech Recognition." 2018 14th IEEE International Conference on Signal Processing (ICSP). IEEE, 2018.
- [11] Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. "A survey of transfer learning." Journal of Big Data 3.1 (2016): 9.
- [12] Song, Guangxiao, et al. "Transfer Learning for Music Genre Classification." International Conference on Intelligence Science. Springer, Cham, 2017.
- [13] Yeom-Song, Víctor, Marisol Zeledón-Córdoba, and Marvin Coto-Jiménez. "A Performance Evaluation of Several Artificial Neural Networks for Mapping Speech Spectrum Parameters." Latin American High Performance Computing Conference. Springer, Cham, 2019.
- [14] Erro, Daniel, et al. "HNM-based MFCC+ F0 extractor applied to statistical speech synthesis." IEEE ICASSP, 2011.
- [15] Greff, Klaus, et al. "LSTM: A search space odyssey." IEEE transactions on neural networks and learning systems 28.10 (2016): 2222–2232.
- [16] Kominek, John, and Alan W. Black. "The CMU Arctic speech databases." Fifth ISCA workshop on speech synthesis. 2004.

GPU based approach for fast generation of robot capability representations

Enfoque basado en GPU para la generación rápida de representaciones de capacidad de robot

Daniel García-Vaglio¹, Federico Ruiz-Ugalde²

Hernández-Zamora, M.F. Gpu based approach for fast generation of robot capability representations. *Tecnología en Marcha*. Tecnología en marcha. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 50-57.

 <https://doi.org/10.18845/tm.v35i8.6449>

1 Electrical Engineering Department. University of Costa Rica. Costa Rica
2 Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: federico.ruizugalde@ucr.ac.cr

Keywords

GPU computation; capability maps; robotic dexterity.

Abstract

Capability maps are an important tool for enabling robots to understand their bodies by providing a way of representing the dexterity of their arms. They are usually treated as static data structures because of how computationally intensive they are to generate. We present a method for generating capability maps taking advantage of the parallelization that modern GPUs offer such that these maps are generated approximately 50 times faster than previous implementations. This system could be used in situations where the robot has to generate these maps fast, for example when using unknown tools.

Palabras clave

Cálculo de GPU; mapas de capacidad; destreza robótica.

Resumen

Los mapas de capacidad son una herramienta importante para permitir que los robots comprendan sus cuerpos al proporcionar una forma de representar la destreza de sus brazos. Por lo general, se tratan como estructuras de datos estáticas debido a la intensidad en computación que pueden generar. Presentamos un método para generar mapas de capacidad aprovechando la paralelización que ofrecen las GPU modernas, de modo que estos mapas se generan aproximadamente 50 veces más rápido que las implementaciones anteriores. Este sistema podría utilizarse en situaciones en las que el robot tiene que generar estos mapas rápidamente, por ejemplo, cuando se utilizan herramientas desconocidas.

Introduction

One of the challenges of modern robotics is to build robots that can work as competent autonomous assistants for humans, such that we can share the same workspaces. It is necessary to enable robots to understand their bodies and their capabilities. This would enable robots to reason about how to do certain tasks and whether or not they need tools or help from other agents to complete them. In particular we are interested on representing the dexterity of a robot, to enable it to eventually reason about the best ways to manipulate objects (find strategies for optimizing dexterity) [10,7].

One approach for representing robotic dexterity are capability maps. These are maps that assign a reachability measure to every point in space for a robotic arm; that is, how many ways the robot is able to reach a certain point in space [8]. Figure 1b shows an example of the capability map of the KUKA LWR 4+ arm (figure 1a). These maps have multiple applications that will be discussed in section 2.

One of the drawbacks of capability maps is that they are very computation-ally intensive to generate, so they are normally calculated once and treated as static data structures [5,2]. But this is not always the most convenient approach. When designing a robotic arm, it is useful to see how changing certain parameters would affect the capabilities of the arm. Having rapid feedback can speed-up the design process. Another example where capability maps could change are if the arm of a robot changes while executing tasks. If a joint gets damaged and the robot decides that the best strategy is to lock it in-place to avoid further damages, the capability

map changes. Also, when the robot is using a tool we are interested on the tip of the tool rather than the tip of the arm. So, whenever the robot picks a new previously unknown tool, it is possible that it will require to generate a new capability map that takes into consideration the effects of the tool.

As shown previously, there are certain situations where we would need to generate capability maps fast. In this paper we present a method for generating them using the parallel computation capabilities of modern GPUs and show how this approach differs from previous implementations.

Related work

Since very early in the development of redundant robotic manipulators, there has been a discussion about how to measure dexterity. One of the proposed dexterity measures is to analyze the amount of orientations that a robot can reach at the point of interest [4]. This idea was then expanded into capability maps, which are maps that represent how many orientations a robotic arm is able reach in every point of its workspace [11]. The points where the robot is able to reach many orientations are considered points of high dexterity, while points that are only reached with one orientation are considered points of low dexterity.

The first generation approaches only took into consideration using inverse kinematics for deciding if a pose was reachable or not [11]. This generation method was very slow. To increase performance a method that used forward kinematics was introduced, but this meant an accuracy penalty [9].

Capability maps have been used in a variety of applications. They have been used in planning, for deciding where to position the body of a robot in front of a table to optimize its manipulation capabilities [6], they have also been used to understand dual-arm manipulation, both for control [5] and for designing hardware [1].

There is a previous example of generating the workspace of a robotic arm with a GPU based approach. They were able to generate the workspace in less than a second [3]. The workspace is the set of positions that can be reached by the robot, but does not encode how well they are reached. They generated random robot poses and incorporated them into a point-cloud to build the workspace. This problem is easier because they don't have to compute how well a point is reached, so they only need to reach each point once, but when generating a capability map each point has to be visited multiple times (from 500 to 6000 in the case of our experiments).

System Overview

The first step is to compute the hierarchical subdivision of space to be able to do the analysis [8]. Let C be the maximum reach length of the arm, we only take into consideration the cube centered on the arm origin of size $2C \times 2C \times 2C$. Each dimension is divided into N_c equal segments, giving a total of N_c^3 voxel. For generating the orientations we used the Euler angles intrinsic convention, also known as yaw, pitch and roll. For generating the pitch and yaw we considered the problem of dividing a sphere's surface into N_s equal areas. This was achieved by creating a Fibonacci spiral where the n_s th section center is given by (1), with φ being the golden ratio, as shown in figure 1c. And finally the roll is generated by dividing 2π into N_r equal parts.

$$\theta_{yaw} = 2\pi (2 - \varphi) n_s, \theta_{pitch} = \arcsin \left(-1 + 2 \frac{n_s}{N_s}\right) \quad (1)$$

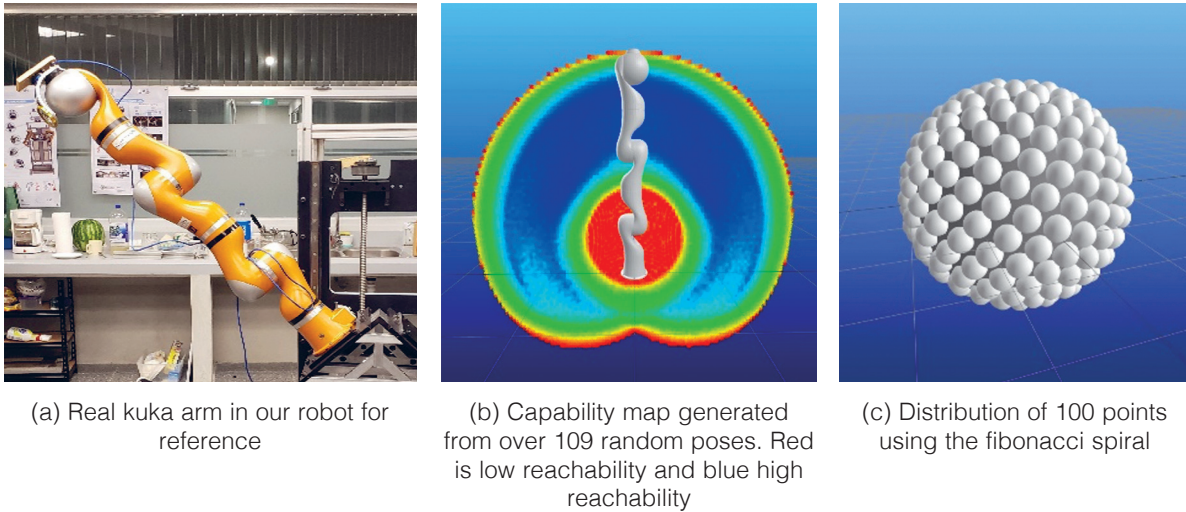


Figure 1. Data generation examples.

There are three techniques to compute the scores for the voxels to create the capability map, use only forward kinematics, only inverse kinematics or the hybrid approach. Because we are seeking a very fast generation of the map, the pure forward kinematics is used. There is an accuracy penalty (11.45% loss in the worst case) that is paid for using only forward kinematics [9].

It is necessary to generate random poses of the end-effector, for this purpose we generate a random number for each joint from a uniform distribution inside the joints limits and it is used as the joint state. Then the forward kinematics algorithm is used to compute the pose of the end effector. These poses are independent from one another, therefore this is done in parallel in the GPU. The parallel forward kinematics solver was implemented from scratch. This is done by generating the homogenous matrices that represent the transformations of each joint and link of the kinematic chain that model the arm and multiplying them in order [9].

The next step is to build the reachability map, this is to find the voxel in which the end-effector is, and which orientation from the generated ones is the closest. Although the reachability map is a 6-dimensional structure (one dimension for each degree of freedom), we find it more efficient to represent it as a 2D matrix called reachability matrix. There is a row for each possible position, and a column for each possible orientation. Then the reachability map is represented by writing a 1 (or a True) in the (c_p, c_θ) coordinates of the matrix (where c_p is the position of the voxel containing the end effector, and c_θ the closest orientation). So, a 1 means that the position and orientation was reached, and a 0 means that it wasn't.

For finding the position and orientation we do not do a normal search, because it is very slow, instead we take advantage of the fact that we know how the voxels and the orientations were generated to find the solution faster. The coordinate inside the reachability matrix for the row (position) is given by (2) where $r(\cdot)$ is the round function, s is the step between voxel centers, (x, y, z) is the exact position of the end-effector, and (x_0, y_0, z_0) is the position of the first voxel.

$$c_p = r\left(\frac{z - z_0}{s}\right) + N_c r\left(\frac{y - y_0}{s}\right) + N_c^2 r\left(\frac{x - x_0}{s}\right) \quad (2)$$

We are using 3 x 3 rotation matrices for representing rotations, because of how well they work with forward kinematics. Let M be the matrix that represents the orientation of the end-effector, we know because of the intrinsic yaw-pitch-roll that the entry $M(2,0)$ represents the sine of the

pitch $-\sin(\theta_{\text{pitch}})$. Then solving for n_s in (1) one knows that C_θ is between C_s and $C_s + N_r$ where $C_s = \lfloor \frac{N_r}{2}(1-M(2,0)) \rfloor$. This means that for finding the nearest orientation we have reduced the search to a list of orientation of size N_r , instead of $N_s \times N_r$.

To find the closest orientation we need to calculate the angle between each orientation and the resulting orientation of the end-effector. Let A and B be two rotation matrices, the angle between them is given by $\arccos((T_r(AB^T)-1)/2)$. The \arccos function is very expensive to compute, so to avoid that we use the fact that it is monotonous in $[0, \pi]$, therefore order relations prevail when using other functions (that are also monotonous). So, instead we calculate $\alpha = 1 - (T_r(AB^T)-1)/2$, and then take the orientation that gives the minimal α . Doing this optimization almost duplicated the speed of the computation.

Filling the reachability matrix is done in parallel for each generated random end-effector pose. The process is completely independent until the reachability matrix has to be written. The advantage is that this is a write-only data structure for this part of the algorithm. The pose generation process and filling the reachability matrix was implemented in a single CUDA kernel of size 1024×1024 (blocks per grid, and threads per block respectively). It gives a total of 1048576 threads, each one with a different random end-effector pose. This are not enough poses to make sure each voxel was visited the right amount of times, so we have to execute the kernel N_B times, where $N_B = \lceil \frac{6}{6250000} N_c^3 N_s N_r \rceil$. There is a linear complexity in regards to the amount of voxels and N_s , but a quadratic complexity in regards to N_r .

The last step is to calculate the capability map from the reachability matrix. This is done by calculating the sum of each row of the reachability matrix. The parallelization of this step is trivial, and it was implemented in a separate kernel.

Once the summation is done there is a score for each voxel. The scores are then and converted into the colour scale (red for 0 and blue for 1).

Results

We ran our system in a GeForce GTX 1060 with 1280 CUDA cores and 6GB of dedicated video memory. It was implemented with the CUDA framework using the numba just in time compiler for Python3.8. The experiments consisted of generating capability maps with different N_c , N_s and N_r values, for the KUKA LWR 4+ robotic arm.

The first results that should be discussed is how the capability map can change while the robot is executing manipulation tasks. Figure 2a shows the capability map of the KUKA LWR 4+ robotic arm if the fourth joint (commonly known as elbow) gets locked at 0° , it is different to the normal capability map presented in figure 1b. Because there is no elbow, the workspace is limited to a hollow shape, almost like the outer layer of the normal capability map. It is clear that if a joint has to be locked the normal capability map must not be used anymore and the robot is required to calculate the new capability map. Figure 2b shows the capability map of the arm if a 30cm tool is attached to the end of the arm, not only the capability map grows in size, but the distribution of high capability regions changes completely. This shows why it is important to generate a new capability map when a new tool is being used. In both cases the capability map must be generated fast so that the robot can continue executing tasks.

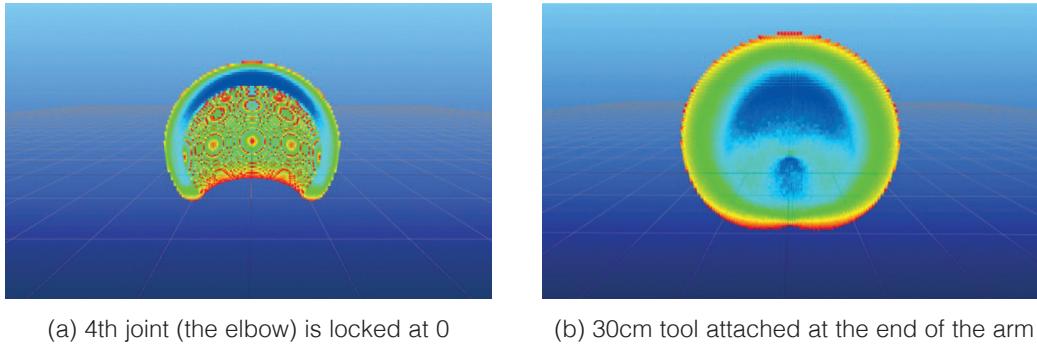


Figure 2. Capability maps of the KUKA LWR 4+ in different kinematic configurations

Although the main objective is to give robots the capacity to quickly generate capability maps during task execution the GPU approach could also be used to generate detailed maps with very high resolution. For this end, it is necessary to analyze how the system behaves as the resolution grows. In figures 3a, 3b and 3c we show how the execution time behaves as the voxel count (N_c^3), N_s and N_r grow. This graphs were generated by leaving the other two parameters constant at $N_c = 50$, $N_s = 50$, and $N_r = 10$. The first two show a linear behaviour while the last one shows a quadratic behaviour as predicted in section 3. The main limitation that is not letting the capability map grow is memory. The reachability matrix has to be stored in GPU memory so that the algorithm can run fast, then the upper bound is the GPU memory, which in our case was 6GB.

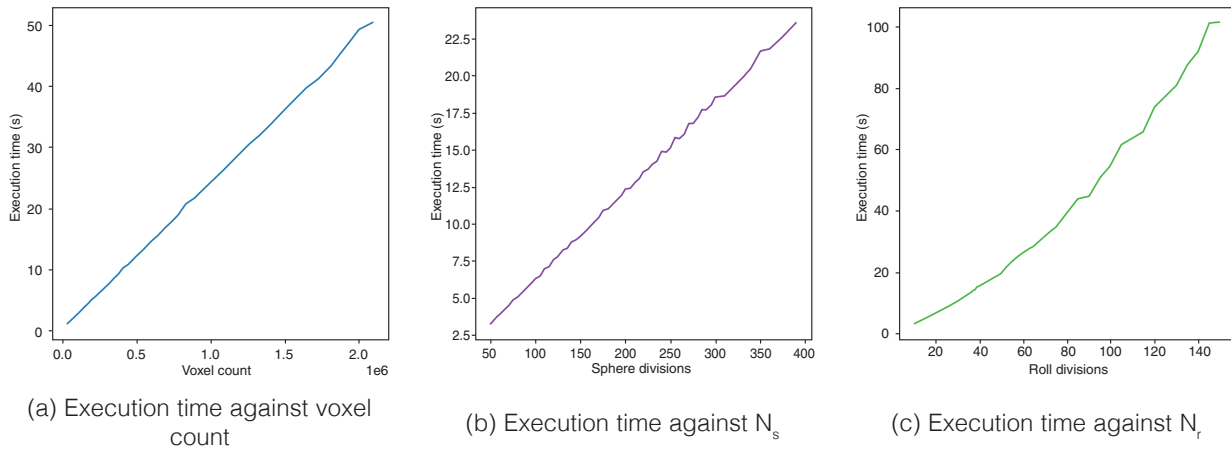


Figure 3. Behaviour of execution time against different workspace subdivision parameters

Finally, our system should be compared to other implementations of capability maps generators. Most literature is focused on using capability maps, but not in generating them. From the papers that discuss the problem of generating them, the authors in [9] explained optimization strategies for generating this structures. They offered the fastest generation times that we could find in the literature, but they were running on CPUs. As we show in table 1 we were able to generate capability maps around 50 times faster once the algorithm was implemented for GPU. We offer a comparison to all the resolution configurations that produce a reachability matrix that fits in our GPU's memory, but in all cases our implementation was considerably faster.

Conclusions and future work

We presented a method for fast generation of capability maps that could be used in situations where the robot needs to calculate them because of changes in the kinematic configuration of its arms. This could be due to damages in a joint that forces it to get locked in-place, or if the robot takes an unknown tool that changes its distribution of dexterity in space. We were able to generate capability maps around 50 times faster than previous implementations which reduced the generation time from the order of minutes to the order of seconds.

Table 1. Comparison between a previous implementation and ours.

N_c	N_s	N_r	Previous	Ours	Speed-up
32	50	10	49.8	1.07	46.54
	100	20	199.8	3.86	51.76
	200	300	552.6	12.02	45.97
47	50	10	139.8	2.71	51.59
	100	20	615.6	10.88	56.58
	200	300	1734.6	36.19	47.65
94	50	100	1264.8	20.47	61.78

This speed increment was achieved because the algorithm was implemented for the GPU instead of the CPU, which is better suited for data parallel problems. Now that vision algorithms are increasingly complex, GPUs are becoming more common in robotic platforms. This enables roboticists to add more algorithms that depend on this hardware like the one presented in this paper. One drawback is that while the capability map is being generated the GPU is utilized at 100% for most of the generation process, so complex vision processes would not be able to run for that period of time.

The biggest limitation is memory. The reachability matrix consumes a lot of space as the resolution (N_s , N_c and N_r) increases. Because the end-effector poses are generated at random from a uniform distribution at the joint-space, it is necessary to have the entire matrix loaded in GPU memory. One part of the future work is to find ways for allowing the reachability matrix grow in a way that performance doesn't get too negatively impacted but better resolutions are also possible.

This project is part of our efforts of building a humanoid robot to assist people (see figure 1a) in real-life scenarios. So, another next step is to incorporate this project into our Humanoid robot cognitive system. This would allow us to test how fast generation of capability maps augment the manipulation capabilities of the robot by enabling it to generate new capability maps “on the fly” for the tools it uses.

Acknowledgment

The present paper is a partial result of a research project entitled “Manipulation of everyday objects using an object model system. Creation of an object model library and a new object model, 322-B6-279” funded by the Research Agency of the University of Costa Rica (Vicerrectoría de Investigación). Additional funding and support is provided by the University of Costa Rica

Dean's office, Electrical Engineering department, Graduate Program in Electrical Engineering, Engineering Research Institute and by the Institute for Artificial Intelligence at the University of Bremen.

References

- [1] Chaves-Arbaiza, I., García-Vaglio, D., Ruiz-Ugalde, F.: Smart Placement of a Two- Arm Assembly for An Everyday Object Manipulation Humanoid Robot Based on Capability Maps. In: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWobi), pp. 1–9 (2018). DOI 10.1109/IWobi.2018.8464192
- [2] Forstnhäusler, M., Wetner, T., Dietmayer, K.: Optimized Mobile Robot Positioning for better Utilization of the Workspace of an attached Manipulator. In: 2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pp. 2074–2079 (2020). DOI 10.1109/AIM43001.2020.9158922. ISSN: 2159-6255
- [3] Jauer, P., Kuhlemann, I., Ernst, F., Schweikard, A.: GPU-based real-time 3D workspace generation of arbitrary serial manipulators. In: 2016 2nd International Conference on Control, Automation and Robotics (ICCAR), pp. 56–61 (2016). DOI 10.1109/ICCAR.2016.7486698
- [4] Klein, C.A., Blaho, B.E.: Dexterity Measures for the Design and Control of Kinematically Redundant Manipulators. *The International Journal of Robotics Research* 6(2), 72–83 (1987). DOI 10.1177/027836498700600206. URL <http://journals.sagepub.com/doi/10.1177/027836498700600206>
- [5] Liu, Q., Chen, C.Y., Wang, C., Wang, W.: Common workspace analysis for a dual-arm robot based on reachability. In: 2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), pp. 797–802 (2017). DOI 10.1109/ICCIS. 2017.8274881. ISSN: 2326-8239
- [6] Makhal, A., Goins, A.K.: Reuleaux: Robot Base Placement by Reachability Analysis. In: 2018 Second IEEE International Conference on Robotic Computing (IRC), pp. 137–142 (2018). DOI 10.1109/IRC.2018.00028
- [7] Morgan, A.S., Hang, K., Bircher, W.G., Alladkani, F.M., Gandhi, A., Calli, B., Dollar, A.M.: Benchmarking Cluttered Robot Pick-and-Place Manipulation With the Box and Blocks Test. *IEEE Robotics and Automation Letters* 5(2), 454–461 (2020). DOI 10.1109/LRA.2019.2961053. Conference Name: IEEE Robotics and Automation Letters
- [8] Porges, O., Lampariello, R., Artigas, J., Wedler, A., Borst, C., Roa, M.A.: Reachability and dexterity: Analysis and applications for space robotics. In: *Proceedings of the Workshop on Advanced Space Technologies for Robotics and Automation (ASTRA)* (2015)
- [9] Porges, O., Stouraitis, T., Borst, C., Roa, M.A.: Reachability and Capability Analysis for Manipulation Tasks. In: M.A. Armada, A. Sanfeliu, M. Ferre (eds.) *ROBOT2013: First Iberian Robotics Conference, Advances in Intelligent Systems and Computing*, pp. 703–718. Springer International Publishing, Cham (2014). DOI 10.1007/978-3-319-03653-3_50
- [10] Ruiz, E., Mayol-Cuevas, W.: Where can i do this? Geometric Affordances from a Single Example with the Interaction Tensor. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 2192–2199. IEEE, Brisbane, QLD (2018). DOI 10.1109/ICRA.2018.8462835. URL <https://ieeexplore.ieee.org/document/8462835/>
- [11] Zacharias, F., Borst, C., Hirzinger, G.: Capturing robot workspace structure: representing robot capabilities. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3229–3236 (2007). DOI 10.1109/IROS.2007.4399105. ISSN: 2153-0866
- [17] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27 (2011). Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [18] Anguita, D., Ghio, A., Ridella, S., Sterpi, D.: K-fold cross validation for error rate estimate in support vector machines. In: R. Stahlbock, S.F. Crone, S. Lessmann (eds.) *Proceedings of the 2009 International Conference on Data Mining, DMIN 2009, July 13-16, 2009, Las Vegas, USA*, pp. 291–297. CSREA Press (2009)

Personalized patient ventilation at large scale: Mass Ventilation System (MVS)

Ventilación personalizada del paciente a gran escala: Mass Ventilation System (MVS)

Ogbolu Melvin Omone¹, Bence Takács², Roland Dóczy³,
Tivadar Garamvölgyi⁴, László Szücs⁵, Péter Galambos⁶,
Tamás Haidegger⁷, Miklós Vincze⁸, Kristóf Papp⁹, Daniel
Drexler¹⁰, György Eigner¹¹, Abdallah Benhamida¹², Ezter
Koroknai¹³, Peter Dombai¹⁴, Miklos Kozlovsky¹⁵

Omone, O.M; Takács, B.; Dóczy, R.; Garamvölgyi, T.; Szücs, L.; Galambos, P.; Haidegger, T.; Vincze, M.; Papp, K.; Drexler, D.; Eigner, G.; Benhamida, A.; Koroknai, E.; Dombai, P.; Kozlovsky, M. Personalized patient ventilation at large scale: mass ventilation system (mvs). *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 58-66.

 <https://doi.org/10.18845/tm.v35i8.6450>

- 1 BioTech Research Center, EKIK, Óbuda University, Hungary. E-mail: ogbolu.melvin@nik.uni-obuda.hu
 <https://orcid.org/0000-0002-4943-8922>
- 2 Antal Bejczy Center for Intelligent Robotics, EKIK, Óbuda University, Hungary. E-mail: b.takacs@irob.uni-obuda.hu
 <https://orcid.org/0000-0003-4262-7461>
- 3 BioTech Research Center, EKIK, Óbuda University, Hungary. E-mail: doczi.roland@nik.uni-obuda.hu
- 4 Antal Bejczy Center for Intelligent Robotics, EKIK, Óbuda University, Hungary. Correo electrónico: tivadar.garamvolgyi@irob.uni-obuda.hu
- 5 Antal Bejczy Center for Intelligent Robotics, EKIK, Óbuda University, Hungary. E-mail: laszlo.szucs@irob.uni-obuda.hu
 <https://orcid.org/0000-0001-8310-0979>
- 6 Antal Bejczy Center for Intelligent Robotics, EKIK, Óbuda University, Hungary. E-mail: peter.galambos@irob.uni-obuda.hu
 <https://orcid.org/0000-0002-2319-0551>
- 7 Antal Bejczy Center for Intelligent Robotics, EKIK, Óbuda University, Hungary. E-mail: tamas.haidegger@irob.uni-obuda.hu
 <https://orcid.org/0000-0003-1402-1139>
- 8 BioTech Research Center, EKIK, Óbuda University, Hungary. E-mail: miklos.vincze@biotech.uni-obuda.hu
 <https://orcid.org/0000-0003-3220-7535>
- 9 BioTech Research Center, EKIK, Óbuda University, Hungary. E-mail: kristof.papp@biotech.uni-obuda.hu
- 10 Physiological Controls Research Center, EKIK, Óbuda University, Hungary. E-mail: daniel.drexler@physcon.uni-obuda.hu
 <https://orcid.org/0000-0001-6655-4354>
- 11 Physiological Controls Research Center, EKIK, Óbuda University, Hungary. E-mail: gyorgy.eigner@physcon.uni-obuda.hu
 <https://orcid.org/0000-0001-8038-2210>
- 12 BioTech Research Center, EKIK, Óbuda University, Hungary. E-mail: benhamida.abdallah@biotech.uni-obuda.hu
- 13 BioTech Research Center, EKIK, Óbuda University, Hungary. E-mail: eszter.koroknai@biotech.uni-obuda.hu
- 14 BioTech Research Center, EKIK, Óbuda University, Hungary. E-mail: peter.dombai@biotech.uni-obuda.hu
 <https://orcid.org/0000-0002-2175-1300>
- 15 John von Neumann Faculty of Informatics, Óbuda University, Hungary. Medical Device Research Group, LPDS, MTA-SZTAKI. E-mail: kozlovsky.miklos@nik.uni-obuda.hu, kozlovsky.miklos@sztaki.hu
 <https://orcid.org/0000-0001-8096-9628>

Keywords

Mass ventilation system; covid-19; medical ventilator system.

Abstract

This paper describes a Mass Ventilation System (MVS) which serves as a medical ventilator system. It can be used to ventilate large number of COVID-19 patients in parallel (5 – 50+) with personalized respiratory parameters. The system has been designed to be medically suitable for both non-invasive and invasive patient ventilation. It protects healthcare workers with its centralized air filtering solution, it increases the effectiveness of the healthcare workers with its networked communication and it can be operated in a temporary emergency hospital setup. In this paper, we describe the basic concept and building blocks of the system.

Palabras clave

Sistema de ventilación masiva; covid-19; sistema de ventilación médica.

Resumen

Este artículo describe un sistema de ventilación masiva (MVS) que sirve como sistema de ventilación médica. Se puede utilizar para ventilar un gran número de pacientes con COVID-19 en paralelo (5 - 50+) con parámetros respiratorios personalizados. El sistema ha sido diseñado para ser médicamente adecuado para ventilación de pacientes invasiva y no invasiva. Protege a los trabajadores de la salud con su solución de filtrado de aire centralizado, aumenta la efectividad de los trabajadores de la salud con su comunicación en red y se puede operar en una configuración de hospital de emergencia temporal. En este artículo, describimos el concepto básico y los componentes básicos del sistema.

Introduction

The Corona-virus Disease 2019 (COVID-19) pandemic across the globe [1] have caused several new diseases which have emerged in different geographical areas, with pathogens including Ebola virus, Zika virus, Nipah virus, and coronaviruses (CoVs causes downturn in the socio-economy [2] and an increasing mortality rate. Most of the patients experience various severity of respiratory problems. COVID-19 is described as the third kind of coronavirus that has emerged as a pandemic in the 21st century affecting the human respiratory system [3]. Based on the information recorded by World Health Organization (WHO), the COVID-19 outbreak started in 2019 in Wuhan, China [1] several new diseases have emerged in different geographical areas, with pathogens including Ebola virus, Zika virus, Nipah virus, and coronaviruses (CoVs, [4]. According to the National Cancer Institute (NCI), COVID-19 is an eminently infectious life-threatening disease and categorized as a respiratory disease. It spreads from one person to another by having a direct or an indirect contact with an infected person. Oftentimes, the patients experience respiration difficulties as the virus spreads around the respiratory system. Small percent of the patients further develops Acute Respiratory Disease Syndrome (ARDS) like symptoms [5]. Hence, the therapy of such patients involves the usage of medical ventilators. The history of medical/mechanical ventilation started in the beginning of the 20th century. Firstly, the short-term ventilation and resuscitation and later, the long-term ventilation solutions have been realized. The development of such medical ventilator devices are still ongoing as new technologies, methods, requirements, and therapies showing up. This paper provides the description of a new concept-based Mass Ventilation System (MVS) that can be used to

ventilate large number of patients in parallel (5 – 50+) with personalized respiratory parameters [6], [7]. This paper is organized as follows. The second chapter gives a brief summary about the problem, following the state of the art. The fourth chapter presents the MVS architecture and its main building blocks. In the conclusion chapter we conclude our work with a brief provisioning of possible future research directions.

Motivations - Problem definition

1. Medical ventilators currently in use are capable of supplying only one person, and each patient must be provided with a separate ventilator. ARDS patients are using the medical ventilators more than one week long in average, so the available quantity runs out fast [8].
2. Exhaled infectious air exits into the common hospital airspace by the respiratory equipment currently in use, whereby doctors and nurses are at increased risk as they work in an air contaminated environment with high concentration of viruses [9], [10].
3. An important factor for setting up mass health camps is to consider which equipment can be used in the absence of hospital infrastructure, especially in places where there are no drainage pipe outlets and where power distribution is limited to each camp bed [11].
4. Medical ventilators are expensive devices, some countries cannot afford proper medical equipment for the entire population in the case of a pandemic [12], [13].
5. As COVID-19 pandemic is increasing the effective patient monitoring/handling and optimized human resource allocation is crucial by the healthcare system [14]. The health status of the patient should be available in real-time for medical professionals, as they do not have time to acquire the data personally during large scale medical crisis situations [15].

State of the Art

A mechanical ventilator is a medical device which is capable of moving air in and out of the patient's lung in a reliable and safe manner. The history of medical/mechanical ventilation begins with the resuscitation devices such as the Pulmotor, which was introduced by the Dräger company (Lübeck, Germany) in 1907 [16]. After the Second World War, a serious polio epidemic broke out and long-term ventilation was urgently needed at large scale. Starting in the 1980s, ventilators were increasingly equipped with electronic components that made ventilation more and more precise [17], [18]. The devices started to display airway pressure, flow of breathing gas, and other vital information on a monitor as ventilation waveforms. Modern measurement and control technology became possible to regulate ventilation parameters more and more precisely. Medical ventilator systems are highly adjustable and complex devices. The medical professional can set many parameters on the device (pressure, volume, respiratory rate, PEEP, oxygen percentage, etc.) according to the predefined treatment. The two major patient ventilation mode is the non-invasive (e.g., using a face mask) and the invasive (using an endotracheal tube) mode [12]. In non-invasive mode the patient can breathe and has consciousness [19]. Regarding modes of ventilation, there are several types in use such as assist-control (AC), synchronized intermittent mechanical ventilation (SIMV), and pressure support ventilation (PSV). The ventilator can then be set to provide a given volume/pressure. In each mode, certain parameters must be set on the ventilator, including respiratory rate (RR), inspiratory flow rate (IFR), the fraction of inspired oxygen (FIO_2), and positive end-expiratory pressure (PEEP). Three main ventilation concepts are available recently (Fig. 1). The single person ventilation is the traditional concept (we will not detail this concept more in this paper), co-ventilation was invented during the COVID-19 pandemic as a fast hacking, problem fixing solution to deal with the huge shortage

of medical ventilators. Mass ventilation setups, as the one proposed here in this paper, are new innovative concepts, which provide cost effective and feasible solutions to the resource shortage problem and has opened a new niche market in the medical ventilator field.




Single person ventilation	Co-ventilation	Mass ventilation
1 patient/device	1-4 patients/device	Many (50+) patients/system
Restricted setup, use dedicated resources, robust and reliable	Reuse free, unallocated capacity, non-reliable, performance at the edge of the capacity	Large scale setup, heavy duty components, robust, reliable use dedicated resources
Individual/personalized ventilation setup	Group ventilation setup	Individual/personalized ventilation setup
No cross-contamination possible	Cross-contamination possible	No cross-contamination possible
		
Analogy: Car/moped/bicycle	Analogy: Car/moped/bicycle sharing	Analogy: Public transport

Figure 1. Overview of a ventilation concepts

A good example of “co-ventilation” was developed by Dr. Alain Gauthier, who has doubled the number of patients per ventilator at his hospital in March 2020 during COVID-19 crisis [20]. A more hybrid co-ventilation solution is the recently developed Individualized System for Augmenting Ventilator Efficacy (iSAVE) system built at MIT, which has overcome some constraints of the co-ventilation method [21].

Mass Ventilation System (MVS) solutions

A fully functional mass ventilation system (the Breathing Aid from Germany) was announced late March 2020 almost immediately after the MassVentil project was officially announced on the Internet [22]. This system contains similar innovations as our KTG-type MVS, but it is only able to do non-invasive ventilation using face mask, in the so called Continuous Positive Airway Pressure (CPAP) mode. Their extended prototypes can ventilate 10 - 30 patients in parallel.

The KTG-type Mass Ventilation System (Mass Ventil - MVS)

The Mass Ventilation System builds up from many components. A KTG-type Mass Ventilation System (MVS) has four essential features [23]:

- Centralized air management unit provides pressurized air for several patients at the same time;
- Patient ventilation parameters (such as oxygen concentration, pressure, respiration rate, PEEP) can be adjusted individually for each patient according to the doctor’s treatment strategy at the bedside of the patients;
- Transporting exhaled infectious air from the common airspace, thereby significantly reducing the risk of infection for all medical workers/staff;
- The system may be installed on an ad-hoc basis in a non-hospital environment.

And the following main technical features [7]:

- Each patient ventilation unit can perform pressure and volume control as required by the treatment.
- All generic ARDS ventilation modes are supported by the system such as: CMV, CPAP, BiPAP, CIMV, APRV, IPPV, PCV.
- PEEP is available till 25 mbar (with fine grain adjustment) for each patient.
- Maximum continuous flow supported by the system is: 120 L/min/patient.
- Patient triggered free breathing support (inhale and exhale) for each patient is supported.
- Possible to cough into the system for each patient.
- O₂-tank or O₂ concentrator can be connected to the system at the patient ventilation units.
- FiO₂ (oxygen level) adjustment is supported from 21%-100% (with fine grain adjustment) for each patient.
- Humidity and temperature are controlled by HME (HME booster is optional but can be used); optionally, temperature can be controlled in the inhale bus too.
- Exhaled air is always carried out from the patient area and filtered 2x with centralized HEPA (ULPA is optional) air filters before releasing into the air.
- Patient interface: Patient ventilation unit can be connected to standard patient ventilation masks (for non-invasive ventilation), using endotracheal tubes (for invasive ventilation) or special Covid-19 masks with tube type connectors, or to nasal prongs. We are about to carry away the infectious air from the patient in order to protect medical professionals, thus patient interfaces with direct exhale valves are not preferred.

In contrast with single person ventilators, the KTG-type MVS consists of a central duct system alongside a personal ventilator unit equipped with a KTG-type valves for each patient (as shown in Fig. 2.). The central inhalation and exhalation duct system supplies air to and collects gases from all the personal ventilator units. In such system both the air management and the data management are centralized to increase effectiveness and resource allocation [6].

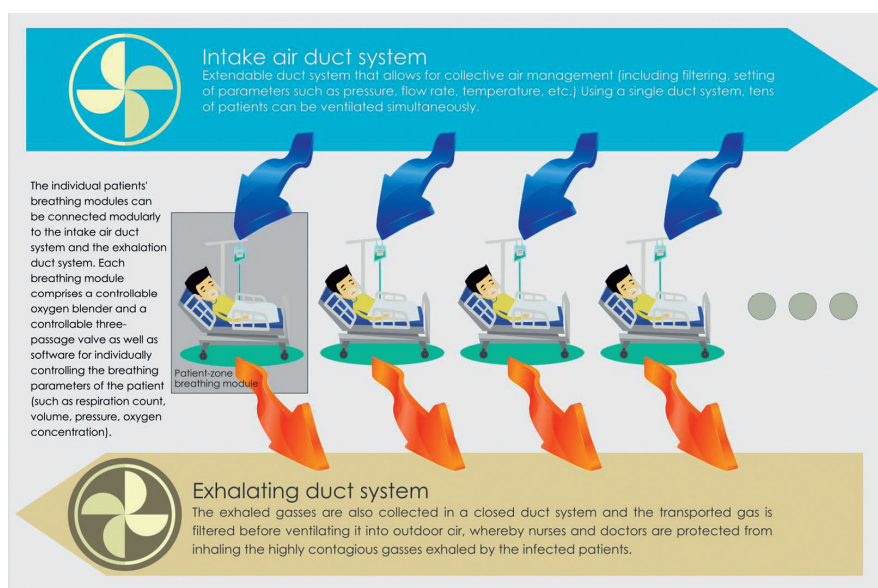


Figure 2. Mass Ventilation System (MVS) design concept

Mass Ventilation System (MVS) Architecture

Figure 3 shows the implemented MVS architecture with its components [23].

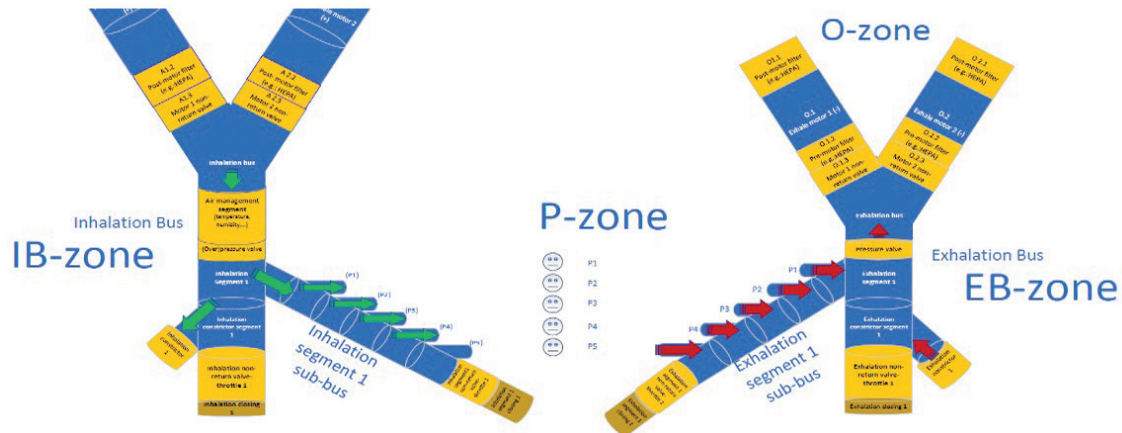


Figure 3. Mass Ventilation System(MVS) architecture [7]

Due to the modularity of the system, new patients can be integrated into the system at the P-zone up to the system's maximum capacity without having to stop ventilating connected patients. The maximum capacity of the system is mainly determined by the diameter of the duct system and the capacity of the motors used in the inhalation and exhalation system. The inhalation air generator system (A-zone) can receive air from different sources. These can be hospital pre-installed aeration tubing, or dedicated ventilator or compressor-based air supply systems. The combination/multiplication of different types of air-generating systems increases the capacity and robustness of the ventilator system (i.e. increasing similar parameters of the entire mass ventilator system). On the inspiratory side (IB-zone), an excess pressure of at least 80 mbar is required. Both the inhalation air supply system and exhalation air dispensing system are redundant in order to remain operational in case of hardware problems and to be able to perform maintenance work (e.g. replacement of filters). A minimum of two fans per inhalation and exhalation subunit is required to achieve adequate redundancy. Fans are fitted with filter(s) to exclude contaminants and pathogens (e.g. COVID-19 virus) of different sizes. Each fan is connected to the duct system via non-return valves (IB-zone and EB-zone). The duct system pressure can be set to the desired value by means of adjustable valves located at the end of the inhalation and exhalation ducts (the main ducts). The flow rate is continuously measured in the main duct to ensure robust operation. In the exhaled air transporting system, there is compression up to the pressure side of the fan, the pressure can be adjusted with a valve at the end of the exhalation duct and it is about 40 mbar lower than the outside air pressure. Due to the lower atmospheric pressure, the pathogens cannot escape from the duct system. The pressure side of the fan is designed in such a way that the air can escape only after appropriate filtering. The duct systems are connected to the personal ventilator unit (as shown below, Fig. 4b.) at the patient's bed via a separate shut-off device, so that new patients can be integrated into the system without any interruption in the whole system. A shut-off fitting is provided at the end of the main ducts for the purpose of scalability. In the personal ventilator unit, a KTG-type valve is controlling the air flow to and from the patient. In the P-zone, the system provides several options for adjusting the O_2 concentration of each patient: oxygen delivery can be achieved by concentration-based or even time-based control. The constant amount of gas delivered per unit time by a mechanical O_2 flow control valve is delivered to the patient by controlling the solenoid

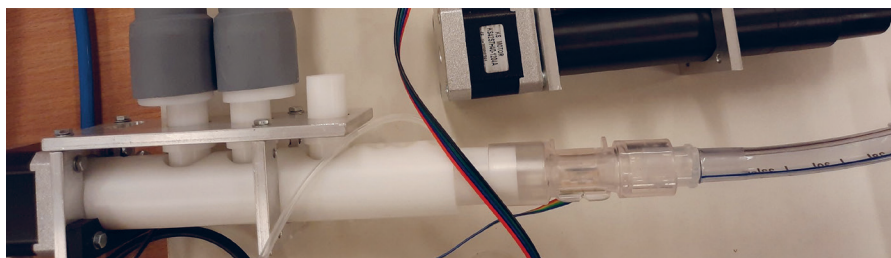
valve at suitable time periods. O₂ gas is delivered to the patient in a separate duct system, thus each patient can have different O₂ level during therapy. The temperature and humidity of the air supplied to the patient is controlled in the P-zone. As in the P-zone humidity and temperature control is easier, thus using passive (e.g. HME – Heat and Moisture Exchangers) or active solutions (e.g. HME Booster) [23]. The whole mass ventilator system and all ventilated patients are monitored by a secure SCADA like unified monitoring system. Measurement data, derivative values and statistics are transmitted to the local server over a communication network using an appropriately encrypted communication channel [7].



(a)



(b)



(c)

Figure 4a. Main air management unit (A and O-zones); 4b. Patient ventilation unit (P-zone); 4c. Patient valve (KTG-type) extracted from the patient ventilation unit (P-zone)

Medical ventilators are classified as a life-critical system - because any mechanical failure may result in death [24], they are highly reliable and carefully designed so that no single point of failure can endanger the patient [25]. The MVS has manual backup mechanisms to enable hand-driven respiration in the absence of power. It has safety valves, which opens to the atmospheric air pressure in the absence of power to act as an anti-suffocation valve for spontaneous breathing

of the patient. It has the possibility to integrate batteries to provide ventilation in case of power failure or defective gas supplies, and methods to operate or call for help if their mechanisms or software fails.

Conclusion

We have designed and realized a Mass Ventilation System (MVS) based on a new so-called mass ventilation concept. The builds up from a centralized air management subsystem, and a patient ventilation unit with a KTG-valve. The system is capable to ventilate many patients in parallel with personalized ventilation parameters. The system supports invasive and non-invasive ventilation modes. The system can save all acquired data from the patient(s) into a database in the background which opens a door into the big data science. The system has been tested successfully in various test environments. The pre-clinical animal trials have been started recently. With the realized MVS, we are planning to ventilate large number of patients in a scalable, safe, reliable and cost-effective way.

Acknowledgment

The authors hereby thank the international MassVentil community and the ITM and NKFIH for their financial support. The research was supported in part by the grant agreement no. 2020-2.1.1-ED-2020-00021 (MassVentil projekt tömeg-lélegeztető rendszer ARDS betegek hatékony ellátásához).

References

- [1] K. Dhama et al., "Coronavirus Disease 2019–COVID-19," *Clin Microbiol Reviews*, vol. 33, no. 4, pp. e00028-20, /cmr/33/4/CMR.00028-20.atom, Jun. 2020.
- [2] "Coronavirus disease COVID-19 pandemic | UNDP." <https://www.undp.org/content/undp/en/home/coronavirus.html>.
- [3] M. F. Bashir et al., "Correlation between environmental pollution indicators and COVID-19 pandemic: A brief study in Californian context," *Environ Res*, vol. 187, p. 109652, Aug. 2020.
- [4] "Archived: WHO Timeline - COVID-19." <https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19>.
- [5] S. K. Gadre et al., "Acute respiratory failure requiring mechanical ventilation in severe chronic obstructive pulmonary disease (COPD):," *Medicine*, vol. 97, no. 17, p. e0487, Apr. 2018.
- [6] "MassVentil Project: Large scale ventilation to beat pandemic." <http://massventil.org/en/massventil-project/>
- [7] "Technical Description - MassVentil." <http://massventil.org/en/technical-description/>
- [8] K. Iyengar, S. Bahl, Raju Vaishya, and A. Vaish, "Challenges and solutions in meeting up the urgent requirement of ventilators for COVID-19 patients," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 499–501, Jul. 2020.
- [9] S. Comunian, D. Dongo, C. Milani, and P. Palestini, "Air Pollution and COVID-19: The Role of Particulate Matter in the Spread and Increase of COVID-19's Morbidity and Mortality," *Int J Environ Res Public Health*, vol. 17, no. 12, Jun. 2020.
- [10] M. Urrutia-Pereira, C. A. Mello-da-Silva, and D. Solé, "COVID-19 and air pollution: A dangerous association?," *Allergol Immunopathol (Madr)*, Jul. 2020.
- [11] G.L. (1994) Historical perspective on the development of mechanical ventilation. In: Tobin M.J. (Hrsg.), *Principles and practice of mechanical ventilation*. 1 – 35.
- [12] L. Brochard, "Mechanical ventilation: invasive versus noninvasive," *European Respiratory Journal*, vol. 22, no. Supplement 47, pp. 31s–37s, Nov. 2003.
- [13] M. M. Feinstein et al., "Considerations for ventilator triage during the COVID-19 pandemic," *Lancet Respir Med*, Apr. 2020.

- [14] D. L. Buckeridge et al., "An infrastructure for real-time population health assessment and monitoring," *IBM J. Res. & Dev.*, vol. 56, no. 5, pp. 2:1-2:11, Sep. 2012.
- [15] Y. Han and H. Yang, "The transmission and diagnosis of 2019 novel coronavirus infection disease (COVID-19): A Chinese perspective," *J Med Virol*, vol. 92, no. 6, pp. 639–644, Jun. 2020.
- [16] "Draeger Pulmotor | Wood Library-Museum." <https://www.woodlibrarymuseum.org/museum/item/96/draeger-pulmotor>
- [17] Division of Pulmonology and Respiratory Intensive Care Unit, San Donato Hospital, Arezzo, Italy et al., "Ventilator Support and Oxygen Therapy in Palliative and End-of-Life Care in the Elderly," *Turk Thorac J*, vol. 21, no. 1, pp. 54–60, Feb. 2020.
- [18] "Ventilators | Clinical Gate." <https://clinicalgate.com/ventilators/> (accessed Aug. 31, 2020).
- [19] "Non-invasive ventilation as long as possible." https://www.draeger.com/en_me/Hospital/Acute-Care-Insights/Respiratory-Care/Mechanical-Ventilation/Prevent.
- [20] "LOOK: Ventilator Hack from Canada - Genius doctor transforms 1 ventilator to 9! - Healthcare Channel COVID-19." <https://healthcarechannel.co/look-ventilator-hack-from-canada-genius-doctor-transforms-1-ventilator-to-9/>
- [21] "About | Individualized System for Augmenting Ventilation Efficacy (iSAVE)." <https://i-save.mit.edu/>
- [22] "Home — Breathing Aid." <https://www.breathing-aid.org/homeen>
- [23] "Downloads - MassVentil." <http://massventil.org/en/downloads>
- [24] Ernst Bahns, "The Breathing-Book: Spontaneous breathing during artificial ventilation", Drager. Technology for Life, pg 1-80.
- [25] Ernst Bahns, "It began with the Pulmotor: The History of Mechanical Ventilation", Drager. Technology for Life, pg 1-114.

Two-dimensional gel electrophoresis image analysis of two *Pseudomonas aeruginosa* clones


Análisis de imágenes bidimensionales de electroforesis en gel de dos clones de *Pseudomonas aeruginosa*

José Arturo Molina-Mora¹, Diana Chinchilla-Montero²,
Carolina Castro-Peña³, Fernando García⁴


Molina-Mora, J.A.; Chinchilla-Montero, D.; Castro-Peña, C.; García, F. Two-dimensional gel electrophoresis image analysis of two *Pseudomonas aeruginosa* clones. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 67-73.

 <https://doi.org/10.18845/tm.v35i8.6452>

1 Centro de Investigación en Enfermedades Tropicales, Universidad de Costa Rica. Facultad de Microbiología, Universidad de Costa Rica. Costa Rica. E-mail: jose.molinamora@ucr.ac.cr

 <https://orcid.org/0000-0001-9764-4192>

2 Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud (INCIENSA). Costa Rica. E-mail: dchinchilla@inciensa.sa.cr

 <https://orcid.org/0000-0002-3093-1346>

3 Centro de Investigación en Enfermedades Tropicales, Universidad de Costa Rica. Facultad de Microbiología, Universidad de Costa Rica. Costa Rica. E-mail: mariacarolina.castro@ucr.ac.cr

4 Centro de Investigación en Enfermedades Tropicales, Universidad de Costa Rica. Facultad de Microbiología, Universidad de Costa Rica. Costa Rica. E-mail: fernando.garcia@ucr.ac.cr

Keywords

2D-GE; Image analysis; CellProfiler; *P. aeruginosa* C25; *P. aeruginosa* C50.

Abstract

A classical strategy to analyse the protein content of a biological sample is the two-dimensional gel electrophoresis (2D-GE). This technique separates proteins by both isoelectric point and molecular weight, and images are taken for subsequent analyses. However, analyses of 2D-GE images require standardized image analysis due to susceptibility of gels to get deformed, presence of overlapping spots and stripes, fuzzy and unstained spots, and others. This represent a difficulty for final users (researchers), which demand for free and user-friendly solutions. We have previously reported the standardization of a protocol to analyse 2D-GE images, and in the current study we applied it to two new bacterial isolates *Pseudomonas aeruginosa* C25 and C50. We first extracted periplasmic proteins after exposure to antibiotics, and we then run a 2D-GE analysis. Images were analysed using our standardized protocol, achieving the identification of protein spots using CellProfiler after pre-processing step. Comparison between strains was done using differential spot analysis, revealing a specific pattern in the protein expression between bacteria. These results will help to study the biological meaning of these strains using proteomic profiling under different conditions.

Palabras clave

2D-GE; análisis de imágenes; CellProfiler; *P. aeruginosa* C25; *P. aeruginosa* C50.

Resumen

Una estrategia clásica para analizar el contenido de proteínas de una muestra biológica es la electroforesis bidimensional en gel (2D-GE). Esta técnica separa las proteínas tanto por punto isoeléctrico como por peso molecular, y se toman imágenes para análisis posteriores. Sin embargo, los análisis de imágenes 2D-GE requieren un análisis de imagen estandarizado debido a la susceptibilidad de los geles a deformarse, la presencia de manchas y rayas superpuestas, manchas borrosas y sin teñir, y otros. Esto representa una dificultad para los usuarios finales (investigadores), que demandan soluciones gratuitas y fáciles de usar. Anteriormente informamos de la estandarización de un protocolo para analizar imágenes 2D-GE, y en el estudio actual lo aplicamos a dos nuevos aislados bacterianos *Pseudomonas aeruginosa* C25 y C50. Primero extrajimos proteínas periplásmicas después de la exposición a antibióticos y luego realizamos un análisis 2D-GE. Las imágenes se analizaron usando nuestro protocolo estandarizado, logrando la identificación de manchas de proteína usando CellProfiler después del paso de preprocesamiento. La comparación entre cepas se realizó mediante análisis de puntos diferenciales, que reveló un patrón específico en la expresión de proteínas entre bacterias. Estos resultados ayudarán a estudiar el significado biológico de estas cepas utilizando perfiles proteómicos en diferentes condiciones.

Introduction

The study of the protein content in biological systems is the main study subject of proteomics. This included not only to identify the particular proteins that are expressed that can explain a biological context, but also the comparison between conditions to recognize differential proteomic patterns [1].

A classical strategy to analyze the proteomic profile of a sample is the two-dimensional gel electrophoresis (2D-GE) [2] this technique is a powerful tool for the analysis and detection of proteins from complex biological sources. Proteins are separated according to isoelectric point by isoelectric focusing in the first dimension, and according to molecular weight by sodium dodecyl sulfate electrophoresis in the second dimension. Since these two parameters are unrelated, it is possible to obtain an almost uniform distribution of protein spots across a two-dimensional gel. This technique has resolved 1100 different components from *Escherichia coli* and should be capable of resolving a maximum of 5000 proteins. A protein containing as little as one disintegration per min of either ¹⁴C or ³⁵S can be detected by autoradiography. A protein which constitutes 10⁻⁴ to 10⁻⁵ of the total protein can be detected and quantified by autoradiography. The reproducibility of the separation is sufficient to permit each spot on one separation to be matched with a spot on a different separation. This technique provides a method for estimation (at the described sensitivities. This technique separates proteins in a layer of polyacrylamide gel by both isoelectric point (pI, pH at which a molecule is electrically neutral) and molecular weight [3], creating spots that are then stained.

Analyses of 2D-GE images require standardized image analysis [3], due to susceptibility of gels to get deformed, presence of overlapping spots and stripes, fuzzy and unstained spots, and others. [1], [4]. However, the 2D-GE image analysis is not straightforward. This represents a difficulty for final users (such as microbiologist, biologist and researchers in general), which demand for user-friendly solutions. However, these user-friendly software are expensive commercial packages. Free options regularly requires command-line work, making it a drawback for researchers.

In this scenario, we have previously reported the standardization of a protocol to analyze 2D-GE images using the Costa Rican bacteria *Pseudomonas aeruginosa* AG1 as model [5]. Now, in this work we applied our protocol to two new isolates, *P. aeruginosa* C25 and C50, which are two clones obtained from the former strain when exposed to high ciprofloxacin (antibiotic) concentrations. *P. aeruginosa* is an opportunistic bacteria able to infect immunocompromised hosts, which is frequently associated with antibiotic multi-resistance [6]. The three Costa Rican isolates have a multi-resistance profile. They are categorized as a high risk clones because are coming from a strain causing infections in hospitals. Thus, the goal of this study was to implement and assess an image analysis protocol using our previously reported protocol to identify protein spots in 2D-GE gels images from two *P. aeruginosa* strains C25 and C50.

To achieve this, we first extracted periplasmic proteins of *P. aeruginosa* C25 and C50 after exposure to antibiotics, and we then run a 2D-GE analysis. Images were analyzed using our standardized protocol, by identifying spots using CellProfiler. Then, comparison between conditions was done using differential spot analysis.

Methods

For the extraction of periplasmic proteins of *P. aeruginosa* C25 and C50, we followed the protocol by [5], [7]. Briefly, cells were cultured until the exponential phase in LB medium. The 2D-GE was performed using strips for separation by isoelectric point (GE HealthCare Immobiline Dry Strip Gels™), and a SDS-GE gradient was done for the molecular weight separations. Images were taken using ChemiDoc™ photo viewer (BioRad®).

The processing step included an image alignment using bUnwarpJ package in the ImageJ program [8] and specifically in biomedical applications that require inter or intra modality image alignment. We have developed an ImageJ plugin called bUnwarpJ, which elastically registers pairs of images. This simple and easy-to-use plugin can be used by researchers and clinicians to create anatomical atlases, segment images using atlases, align pairs of images

distorted by both physical and acquisition related distortions, etc. Registering two images consists on finding the image transformation that maps corresponding pairs of pixels between the original and "distorted" images. We use here the term distorted in a wide sense, to account not only for "sensu strictu" image distortions, but also for anatomical variations between or within individuals. We use an algorithm that simultaneously calculates the direct and inverse transformations and minimizes the similarity error between the target and source images after imposing a consistency constraint. This approach provides bidirectional registration from image A to B or from B to A in a single computation. We use B-splines to represent both images and deformations and make use of a powerful optimizer to converge fast to the best image alignment. Our plugin allows guiding the registration process using the image similarity, the consistency of the deformations, vector-spline regularization and/or a set of optional landmarks, which can be calculated and fed from other ImageJ plugins such as the automatic extractors Scale-Invariant Feature Transform (SIFT). In this program, five spots were used as reference for the deformation of images and to achieve the alignment. Identification of spots was done using our previously reported protocol [5]. Briefly, CellProfiler (<https://CellProfiler.org/>) was used to analyze images following the next steps: images inversion, primary object recognition and segmentation, manual editing, intensity measuring and visualization of objects.

To compare 2D-GE images, a differential spot analysis was implemented. Pairs of images were compared to identify shared spots using an analysis of primary objects (segmentation) of overlapping spots, identification of exclusive spots in each image using the no-overlapping regions, and the subsequent representation spot borders separating shared (red circles) or exclusive dots (green or blue circles).

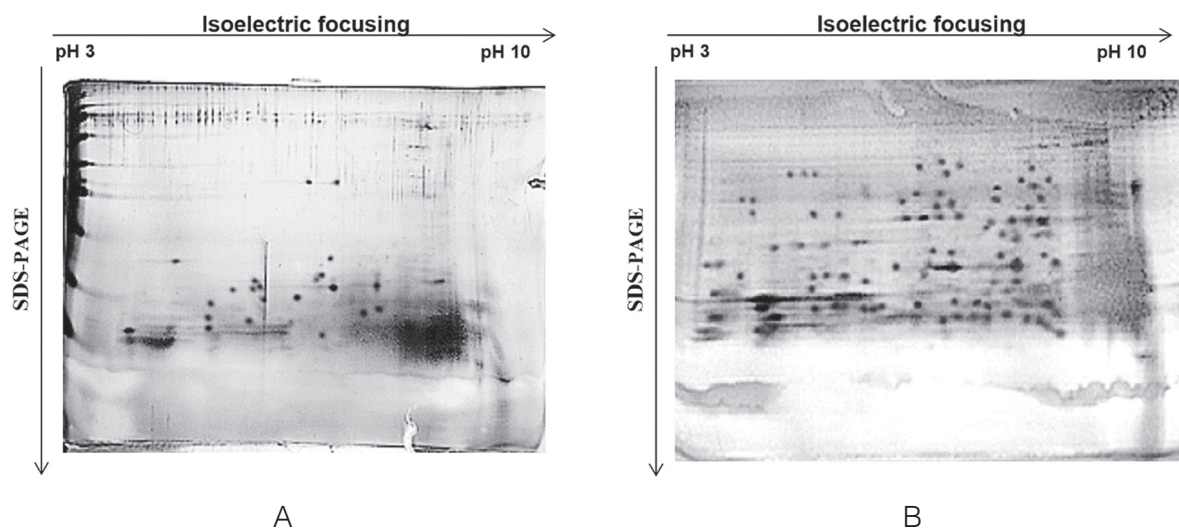


Figure 1. Example of two-dimensional gel electrophoresis (2D-GE) of *P. aeruginosa* C25 (A) and C50 (B) after growing in LB medium. Assays was performed after cells were growth in LB medium.

Results and discussion

Proteomics is considered an essential field for the systematic analysis of biological systems, an assessment of changes in the abundance of proteins that occur in living organisms and that can be studied at various levels [4].

The two-dimensional gel electrophoresis 2D-GE is a classical technique used to analyze the protein content in biological samples [1]. Here we first performed a 2D-GE assay for the bacterial clones *P. aeruginosa* C25 and C50, as shown in Fig. 1-A-B.

However, 2D-GE image analysis requires specific protocols due to image complexity [3]. In this way, we previously established a standardized protocol to identify protein spots using CellProfiler and other image analysis tools [5].

For the pre-processing step, bUnwarpJ package in the ImageJ program was used to align images. According to this pipelines, five points between the target image (to be modified) and a reference image are selected as common denominator to make the alignment, creating a deformation field and grid (Fig. 2-A-B).

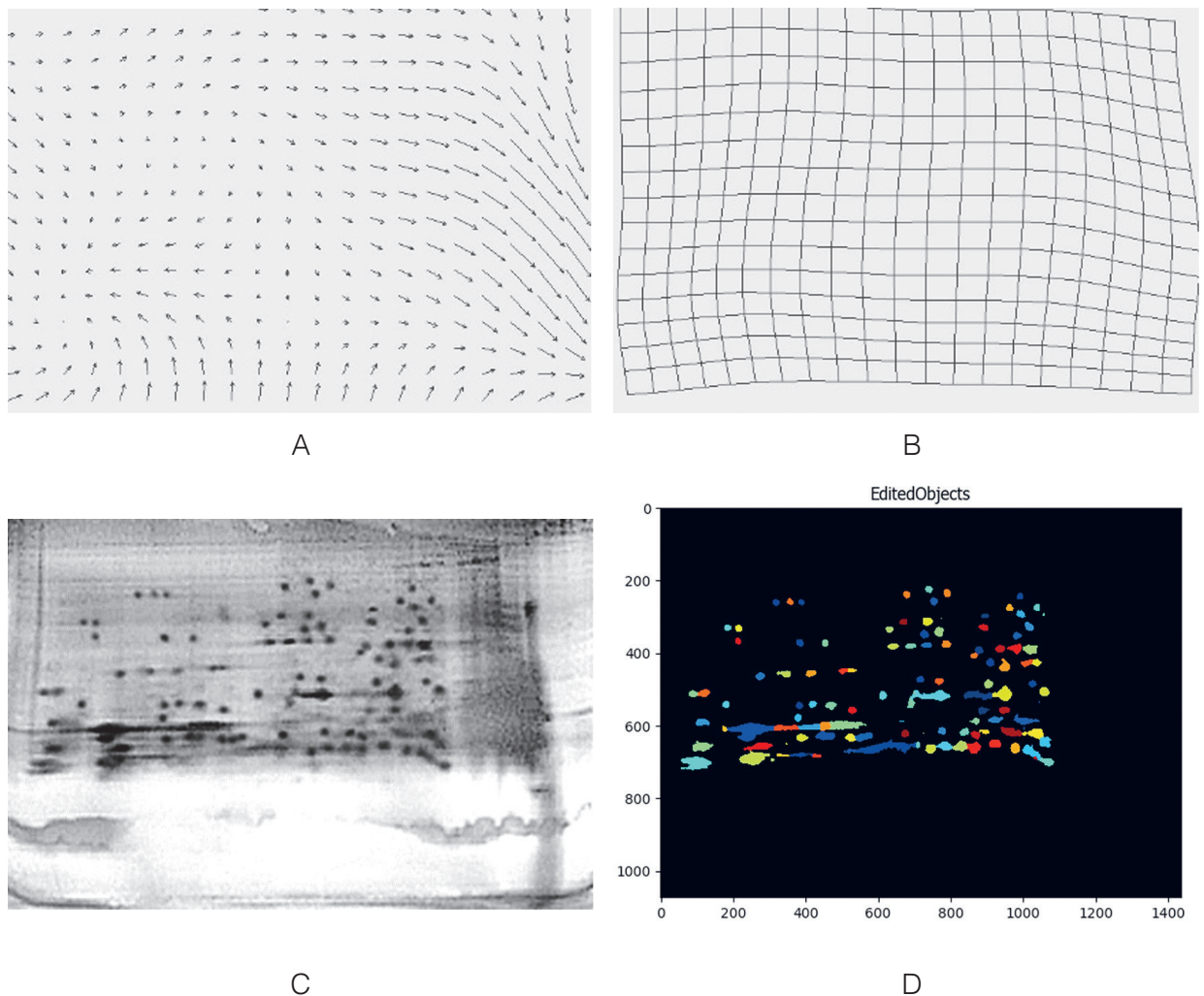


Figure 2. Analysis of 2D-GE images. Examples of deformation field (A) and deformation grid (B) to align images against a reference in the pre-processing step. (C) Example of a raw image used for the identification of spots using CellProfiler pipeline, as resulted in (D).

As shown in Fig. 2-C-D, identification of spots was achieved using CellProfiler software. Different metrics were used to optimize the segmentation algorithm, as previously described [5]. Although automatic spot recognition is sensitive to complex regions, manual edition helped to solve these drawbacks. Commercial solutions have similar tools to deal with this particular features that are common un 2D-GE image analyses [3].

With a modified protocol, the pipeline was also able to recognize common and shared spots when comparison of proteomic profiles of the two strains was done.

For this, a new consensus image was built using image operations (pixel operations), making possible the identification of common spots, which were identified in the same way as before but using the new image. After subtraction of shared dots, exclusive spots were marked and a final visualization was done in the initial images, as shown in Fig. 3.

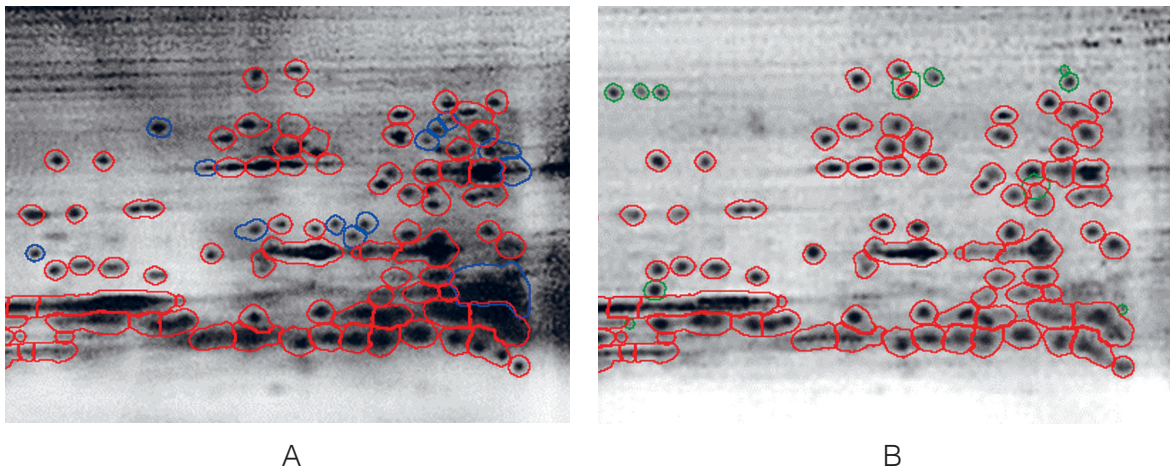


Figure 3. Example of the differential spot identification with 2D-GE images from two *P. aeruginosa* strains C25 (A) and C50 (B). Shared spots were identified using red circles, and exclusive spots were marked as blue or green spots.

Regarding the CellProfiler program, this is a known tool used for cell imaging, for example for microscopy images. However, as we have demonstrated before [5] and here, it is possible to use the algorithms to recognize spots in 2D-GE images. See our previous work for details of the implementations, more details of the pipeline and comparison of samples [5].

In summary, in this work we presented a new analysis of 2D-GE images using a standardized protocol to identify spots and compare conditions by proteomic profile. This was done using two *P. aeruginosa* clones, in which was possible to identify both shared and exclusive dots. Although this work is focused on the image analysis, these results will help us to apply this protocol to study *P. aeruginosa* strains under different experimental conditions, including antibiotics or other stressors and their effect on the proteomic profile of the bacteria.

References

- [1] M. M. Goez, M. C. Torres-Madroñero, S. Röthlisberger, and E. Delgado-Trejos, "Preprocessing of 2-Dimensional Gel Electrophoresis Images Applied to Proteomic Analysis: A Review.," *Genomics. Proteomics Bioinformatics*, vol. 16, no. 1, pp. 63–72, 2018.
- [2] P. H. O'Farrell, "High resolution two-dimensional electrophoresis of proteins.," *J. Biol. Chem.*, vol. 250, no. 10, pp. 4007–21, May 1975.
- [3] M. Natale, B. Maresca, P. Abrescia, and E. M. Bucci, "Image analysis workflow for 2-D electrophoresis gels based on imageJ," *Proteomics Insights*, vol. 4, pp. 37–49, 2011.
- [4] T. S. Silva, N. Richard, J. P. Dias, and P. M. Rodrigues, "Data visualization and feature selection methods in gel-based proteomics.," *Curr. Protein Pept. Sci.*, vol. 15, no. 1, pp. 4–22, Feb. 2014.
- [5] J. A. Molina-Mora, D. Chinchilla-Montero, C. Castro-Peña, and F. Garcia, "Two-dimensional gel electrophoresis (2D-GE) image analysis based on CellProfiler," *Medicine.*, vol. IN-PRESS, 2020.
- [6] R. T. Cirz, B. M. O'Neill, J. A. Hammond, S. R. Head, and F. E. Romesberg, "Defining the *Pseudomonas aeruginosa* SOS response and its role in the global response to the antibiotic ciprofloxacin," *J. Bacteriol.*, vol. 188, no. 20, pp. 7101–7110, Oct. 2006.
- [7] G. F. Ames, C. Prody, and S. Kustu, "Simple, rapid, and quantitative release of periplasmic proteins by chloroform.," *J. Bacteriol.*, vol. 160, no. 3, pp. 1181–3, Dec. 1984.
- [8] I. Arganda-Carreras, C. O. S. Sorzano, J. Kybic, and C. Ortiz-de-solorzano, "bUnwarpJ : Consistent and Elastic Registration in ImageJ . Methods and Applications .," *Image (Rochester, N.Y.)*, 2006.


Assessing costa rican children speech recognition by humans and machines


Evaluación del reconocimiento de voz de los niños costarricenses por humanos y máquinas

Maribel Morales-Rodríguez¹, Marvin Coto-Jiménez²

Morales-Rodríguez, M.; Coto-Jiménez, M. E. Assessing costa rican children speech recognition by humans and machines. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 74-82.

 <https://doi.org/10.18845/tm.v35i8.6453>

1 University of Costa Rica. Costa Rica.
E-mail: maribel.moralesrodriguez@ucr.ac.cr
 <https://orcid.org/0000-0002-3426-5192>

2 University of Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr
 <https://orcid.org/0000-0002-6833-9938>

Keywords

Children speech; speech recognition; speech technologies; WER.

Abstract

In recent years, an increasing number of studies on human-computer interaction is taking place, due to the pervasive speech interfaces implemented in systems such as cell phones, personal and home automation assistants. These studies include automatic speech recognition (ASR) and speech synthesis, and are considering a wider variety of conditions of the signals, such as noise and reverberation, and accents and age-related effects as well. For example, one of the key challenges is the development of ASR for children's speech. Since the current systems have a dependency on language and accents, thus, to improve it, the investigations of speech recognition technologies suitable for children are needed. In this paper, we assess commercial ASR systems for the recognition of Costa Rican children's speech, for users with ages ranging between three and fourteen years old. To establish a comparison and numeric validation of the ASR systems in recognizing children's isolated words, we conducted a large subjective listening test that computes the differences and challenges that remains for the state-of-the art ASR systems. The results provide evident numeric differences between ASR systems and human perceptions, especially for younger children. Additionally, we provide suggestions for future research directions in the field.

Palabras clave

Habla de niños; reconocimiento de voz; tecnologías del habla; WER.

Resumen

En los últimos años, se está llevando a cabo un número creciente de estudios sobre la interacción persona-computadora, debido a las interfaces de habla generalizadas implementadas en sistemas como teléfonos celulares, asistentes personales y de automatización del hogar. Estos estudios incluyen el reconocimiento automático del habla (ASR) y la síntesis del habla, y están considerando una variedad más amplia de condiciones de las señales, como el ruido y la reverberación, y también los acentos y los efectos relacionados con la edad. Por ejemplo, uno de los desafíos clave es el desarrollo de ASR para el habla de los niños. Dado que los sistemas actuales tienen una dependencia del lenguaje y los acentos, por lo tanto, para mejorarlo, se necesitan las investigaciones de tecnologías de reconocimiento de voz adecuadas para los niños. En este trabajo evaluamos sistemas ASR comerciales para el reconocimiento del habla infantil costarricense, para usuarios con edades comprendidas entre los tres y los catorce años. Para establecer una comparación y validación numérica de los sistemas ASR para reconocer las palabras aisladas de los niños, realizamos una gran prueba de comprensión auditiva subjetiva que calcula las diferencias y desafíos que quedan para los sistemas ASR de última generación. Los resultados proporcionan diferencias numéricas evidentes entre los sistemas ASR y las percepciones humanas, especialmente para los niños más pequeños. Además, ofrecemos sugerencias para futuras direcciones de investigación en el campo.

Introduction

During the last decade, significant progress in the field of automatic speech recognition (ASR) for general purpose devices and situations have been built and deployed, including commercial and daily-use applications.

Most of the research and implementation has been devoted to developing systems targeting adult specific speakers [1]. For children, the first studies raised attention to the poor performance of the ASR system for this population [2, 3], so increasing attention has been paid to improve this performance.

The vast majority of the research on children's speech recognition has been made for the English language, with some exceptions on other languages [4, 5] or English as a second language [6, 7].

Among the reasons for the decrease in the effectiveness of the ASR systems on children's speech, can be explained in terms of acoustic features such as higher pitch and formant frequencies, longer segmental duration, and greater temporal and spectral variability [8]. Most ASR systems based on Hidden Markov Models (HMM) or Deep Learning requires a large amount of training data, which is available recording new material that covers all the phonemes, special keywords and vast vocabularies.

These materials are more readily available or produced for adults. For that reason, general purpose ASR systems trained with specific data of adults can be affected by the spectral and temporal variability that characterize the developmental changes in speech production of children.

The pursuit of better systems is motivated by the tremendous potential in children's education, with a wide variety of possible applications ranging from pronunciation training applications to educational games [5]. For example, child-robot interaction is an area with potential contributions to domains such as health-care, education and entertainment [9]. This interaction is expected to occur in the most natural form of communication for humans that is listening, understanding and speaking.

According to [1], word recognition errors of ASR systems can be 100% worse for children's speech particularly at early childhood. The temporal and spectral characteristics that affect children's speech recognition can be aggravated by variabilities introduced by accents or regional changes in speaking styles.

For these reasons, all current ASR systems and the development of speech synthesis or other speech technologies have a dependency on language and accents. In particular, in the case of Costa Rican Spanish, there is little work published in terms of the performance of ASR systems, and even less for children's speech from this population.

In this work, we conducted an exploratory study on the performance of several commercially available ASR systems in recognizing Costa Rican children's speech. These systems can be considered state-of-the art. We also present a comparison with subjective listening tests to provide numerical differences between humans and machines. We report the recording methodology for the children and the procedures to validate the results.

Related work

The effects on variability in children's speech have been studied in [5], for Portuguese children between 3 and 10 years old. The correlation of some characteristics of speech production, such as the truncation of consonant clusters, disfluencies and pronunciation quality, with ASR performance has been observed. For ages 3 to 6 years old, the recognition errors were as high as 69.9%. For children of age 6, recognition errors were found as high as 80% in some conditions [10].

For the English language, where the vast majority of studies have been made, [3] found recognition errors between 21% to 65% in ASR systems. It is clear that such error rates are unacceptable for general purpose applications or for children-computer interaction, such as the one proposed in [11], consisting

of a conversational Nao robot adapted for children's interaction. Some of the relevant features in such interactions include the ability to adapt to an individual child's language proficiency, respond contingently, both temporally and semantically, provide effective feedback and monitor children's learning progress [12]. The necessity for developing speech interfaces for more languages, which consists of ASR system, speech synthesis, natural language processing and other related technologies, has been mentioned in [13], where for most languages sufficiently large corpora of children's speech are often not available for developing speech-driven educational applications. Children would benefit from speech interfaces by using computers, tablets and cell phones, even in ages below which they have acquired reading and writing skills [14]. It is, therefore, of great interest to extend the capability of ASR systems to this speaker category.

When a proper database for developing the recognition of isolated words or sounds for children can be developed, previous studies have shown that a similar accuracy in phone recognition can be achieved with children than adults, despite the higher variability in the children's speech [15].

The rest of this paper is organized as follows: Section 2 presents the methodology we used to collect and analyze data. Also, Section 3 presents the results and comparison between human and machine word error rates, and finally Section 4 presents the conclusions.

Methodology

This work contemplated the recording sessions with children, the transcription and the edition of those recordings. Additionally it contemplate the evaluation of the results, both with ASR and humans. The following subsections give details on these aspects:

Speech material

For the design of the database, an interaction strategy was developed with the participating children in which they used formal and non-formal instruments of oral language assessment. The objective of this interaction was collecting isolated or two-word phrases, instead of assessing the articulation of each phoneme, which was the primary purpose of the instrument used.

The recordings contemplate words by semantic groups of high use in children's language, both in activities of daily life and within the initial school curriculum (colours, animals, food). The recordings were carried out in a professional studio.

Among the strategies used for taking voice samples when interacting with children, it is important to consider that knowledge in the area of child development is required. Not only because of the way in which it is relevant to interact with the participants but also from the aspects of perception and typical attention accorded to age in order to take full advantage of the recordings that need to be collected.

Some of the strategies employed were: the use of material with good visual contrast, utilizing game skills for word pronunciation, alternating data collection with spaces of non-formal interaction or breaks, and the use of verbal positive reinforcement.

Speaker selection

In order to provide the inputs for the children's voice database, several recording sessions were made with children between 3 and 14 years old during this study, both male and female, and all from the central region of Costa Rica.

We split the speakers into three age groups: early childhood (ages 3 to 7), middle childhood (ages 8 to 11 years) and late childhood or early puberty (ages 12 to 14 years). We analyzed the speech recognition according to the age of the participant and also for each group. Due to the difficulties that arose in the verbal interaction with children at its earliest age, 60 words or short phrases were selected for each group.

WER Calculation

To measure the impact on speech recognition, we implemented the Word Error Rate (WER) measure, defined as:

$$\text{WER} = \frac{(D+S+I)}{N} \quad (1)$$

where D is the number of words deleted, S the number of substitutions and I is the number of words inserted. N is the total number of words.

For ASR systems, the measure was implemented in Python, and for human listeners, the calculation was done manually.

Automatic Speech Recognition

For the evaluation of machine ASR, we selected four state-of-the art commercial systems, which include Google Speech Recognition, IBM Watson, Happy Scribe and Cobalt Speech Cubic. We present the result without specifying which system obtained particular WER, because we pretend to assess general machine speech recognition instead of comparing commercial systems.

Each of the ASR systems is capable to personalize and boost results in several ways. For example, keywords or expressions to spot, model selection (voice commands, video transcriptions, etc.), speaker diarization and specific accent among Spanish variants.

In the first evaluation for this population, we applied each ASR with its default settings, and did not select any of the capabilities that can possibly decrease the error rate, with the exception of the Costa Rican Spanish accent when available.

The children's recordings were presented to each ASR system with two seconds of silence between each word or short phrase. This evaluation of isolated words can represent a disadvantage in terms of the language modeling that increases the ASR's capability when the context, previous and following words are detected and used to transcribe each word inside longer sentences better.

Human evaluation

The human children's speech recognition evaluation involved 42 listeners aged between 20 and 37 years old. Each listener was a native Costa Rican Spanish speaker and lived in the same area as the children that were recorded. The listening sessions were conducted using a simple program running on a computer and a set of quality speakers.

Each participant wrote a transcription of each word by hand, for a total of 180 recordings and more than 7000 results to organize and manually process. The words were presented randomly within the recordings of each group, in the same manner that was presented to the evaluation of ASR by machines.

At the end of the session each transcription was compared to the real words that the children intend to pronounce and the results were organized according to age group.

Results and discussion

The WER results are presented and compared in Figure 1 for children between 3 to 7 years, Figure 2a for years 8 to 11 and Figure 2b for years 12 to 14. It is clear for all cases how human listeners' errors are lower than those of the ASR systems.

The differences increase as the age group increase. It is particularly remarkable how difficult is for computer ASR systems to recognize the words pronounced by children at the earliest age, with a range of 38.81% to 95.52%. This second percentage means that almost all of the isolated words and short phrases pronounced by children are not properly recognized. Virtual assistants, communicative robots, speech to speech translation and any other technology that relies on speech recognition can be seriously limited for this population. On the other hand, the 12.9% of WER evaluated in humans means that human listeners understood most of the words.

The differences are still high for children between the range of 8 to 11 years old. Both human and machine recognition increases its effectiveness for this second age group. With an average of 49.17%, most of the systems with automatic speech recognition will have problems understanding almost half of the messages from children in this age range, where human listeners have only 2.8% of WER. It is vital to consider also that the children's voices were recorded in the ideal, noise-controlled conditions of a professional studio.

For children aged 12 to 14, the average WER decreases to 35.89%. As expected, both for humans and machines, the WER consistently decreases as the age increases, and can be as low as 14.52% for ASR in machines in this age group.

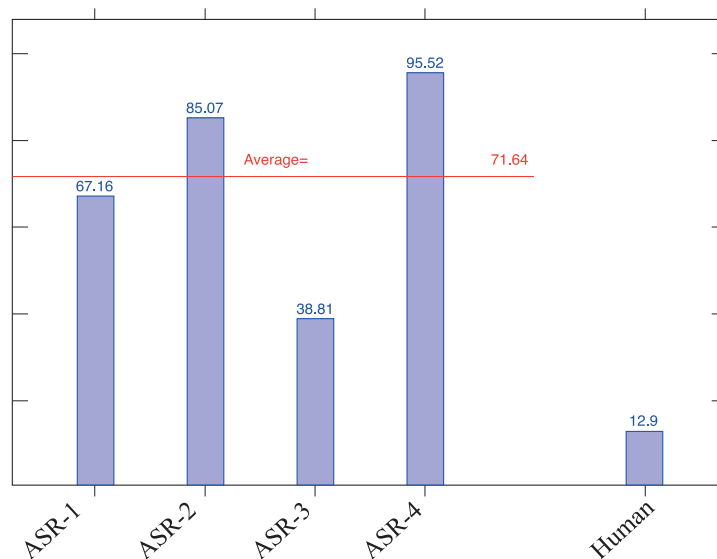


Figure 1. WER results for years 3 to 7

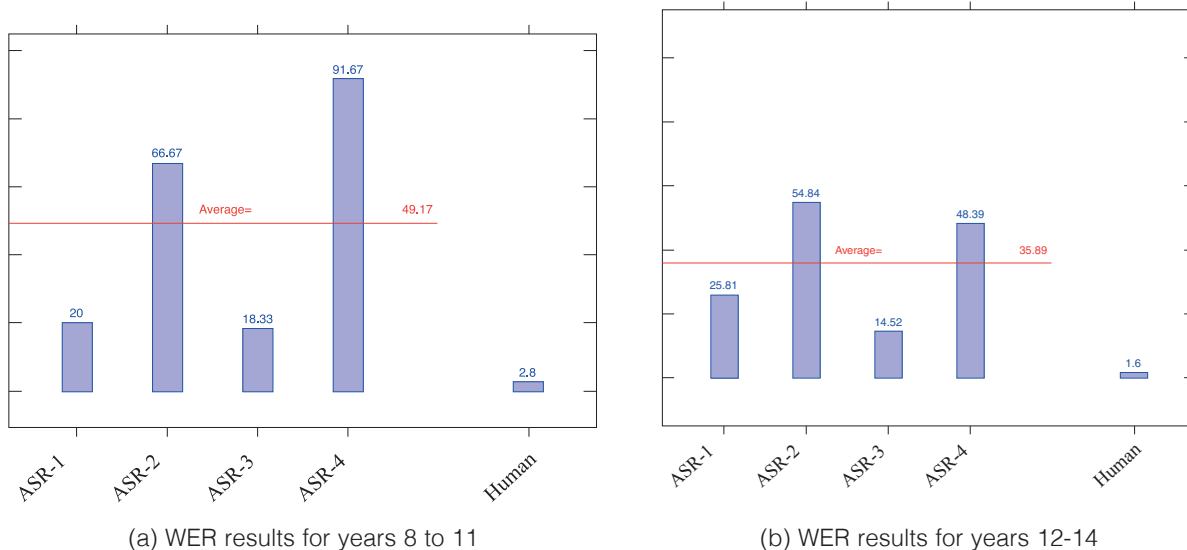


Figure 2. WER results for years 8 to 14

This group is also the one where the ASR systems perform more similarly than the previous groups.

In Figures 3a and 3b human and machine recognition performance is compared, according to gender. In all machine evaluation, the trend of recognition capability is the same for all systems, with similar variability for both female and male children. In general, the female Costa Rican voices present higher errors in early and late childhood when compared to their male counterparts. In comparison to the machine recognition, this trend is not present in the human listeners, where there are similar level or recognition errors at both ends of the plot, and the variance is very low for all the age range.

The errors produced by the human listeners can be explained by the oral production of the children, for which the acquisition of oral language and of course in its development process. For example, the 3 to 4 years old participants are within the stage of linguistic development and the articulation skills in terms

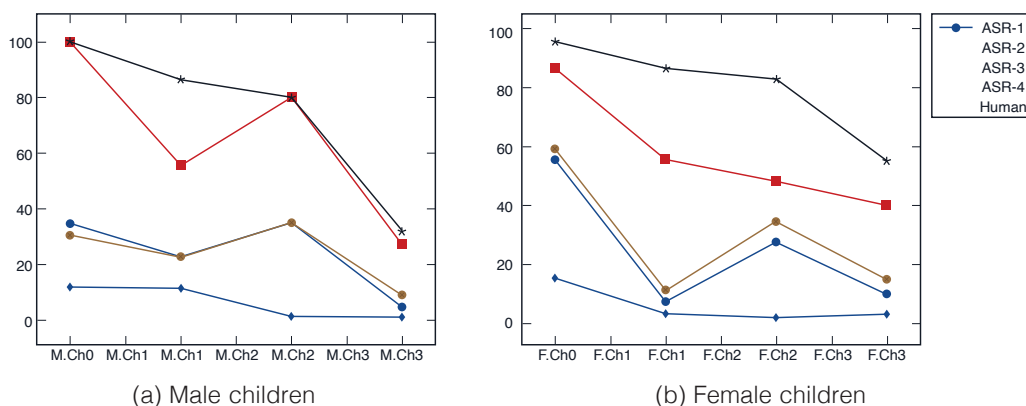


Figure 3. WER according to age and gender

of acquisition of phonemes of both are located in the first stages of development, where it is typically the processes of phonological simplification [16] where the words are simplified.

In the case of 7-year-old children the acquisition of phonemes also known as phonetic-phonological development is an average of 90 % of acquisition in the lateral articulatory modes of the “l” the vibrating modes of the “r” in syllables Fricatives “Pr” and “Fr” as well as “rr” and 80 % in the use of “ll” [17]. This could lead us to suppose that the minor participants phonological development could confuse the human participants and even more the machine recognizers.

Conclusions

In this work we performed a comparison between humans and machines for automatic speech recognition of Costa Rican children ages 3 to 14 years old. For machine recognition we use several recognized commercial systems, and for the human assessment, we conducted a large listening test.

Results show that there is still a big difference between what machines can do in this task, in comparison with humans. Particularly for children in their early years, the machine errors can be as high as 95%, which can render a system to be of no use for children/machine interaction using voice.

As age increases, the machines tend to perform much better in understanding the words in this accent. Still, there are significant differences with human perceptions, that can be as high as 50% in more transcription errors.

This information is useful to assess those differences numerically and it can lead to new research in improving the results with modifications or adaptations of the voices in the systems.

Future work includes the development of systems trained with Costa Rican children’s voices that can be competitive or perform better for automatic recognition in this population.

References

- [1] Gerosa, Matteo, et al. “A review of ASR technologies for children’s speech”. Proceedings of the 2nd Workshop on Child, Computer and Interaction. 2009.
- [2] Russell, Martin, Shona D’Arcy, and Lit Ping Wong. “Recognition of read and spontaneous children’s speech using two new corpora”. Eighth International Conference on Spoken Language Processing. 2004.
- [3] Li, Qun, and Martin J. Russell. “An analysis of the causes of increased error rates in children’s speech recognition”. Seventh International Conference on Spoken Language Processing. 2002.
- [4] Cossi, Piero, et al. “Comparing open source ASR toolkits on Italian children speech”. WOCCI. 2014.
- [5] Hämalainen, Annika, et al. “Correlating ASR errors with developmental changes in speech production: A study of 3-10-year-old European Portuguese children’s speech”. 2014.
- [6] Adi, Derry Pramono, Agustinus Bimo Gumelar, and Ralin Pramasuri Arta Meisa. “Interlanguage of Automatic Speech Recognition. “2019 International Seminar on Application for Technology of Information and Communication (iSemantic). IEEE, 2019.
- [7] Moussalli, Souheila, and Walcir Cardoso. “Intelligent personal assistants: can they understand and be understood by accented L2 learners?”. Computer Assisted Language Learning (2019): 1-26.
- [8] Lee, Sungbok, Alexandros Potamianos, and Shrikanth Narayanan. “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters”. The Journal of the Acoustical Society of America 105.3 (1999): 1455-1468.
- [9] Kennedy, James, et al. “Child speech recognition in human-robot interaction: evaluations and recommendations”. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. 2017.



- [10] D'Arcy, Shona, and Martin Russell. "A comparison of human and computer recognition accuracy for children's speech". Ninth European Conference on Speech Communication and Technology. 2005.
- [11] Kruijff-Korbayov'a, Ivana, et al. "Spoken language processing in a conversational system for child-robot interaction". Third Workshop on Child, Computer and Interaction. 2012.
- [12] Vogt, Paul, et al. "Child-robot interactions for second language tutoring to preschool children". *Frontiers in human neuroscience* 11 (2017): 73.
- [13] Hämalainen, Annika, et al. "A multimodal educational game for 3-10-year-old children: collecting and automatically recognising european portuguese children's speech". *Speech and Language Technology in Education*. 2013.
- [14] Elenius, Daniel, and Mats Blomberg. "Comparing speech recognition for adults and children". *Proceedings of FONETIK 2004* (2004): 156-159.
- [15] Giuliani, Diego, and Matteo Gerosa. "Investigating recognition of children's speech". 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. *Proceedings.(ICASSP'03)*. Vol. 2. IEEE, 2003.
- [16] González, M. J. *Trastornos fonológicos. Teoría y Práctica*. Universidad de Málaga: Secretariado de publicaciones. España, 1989.
- [17] Ortiz Rubia, V. *Procesos fonológicos de simplificación*. Mendoza, Universidad del Aconcagua. Facultad de Ciencias Médicas, 2007. <http://bibliotecadigital.uda.edu.ar/229>.

Colorectal cancer vaccines: in silico identification of tumor-specific antigens associated with frequent HLA-I alleles in the costarican Central Valley population

Vacunas contra el cáncer colorrectal: identificación en silico de antígenos específicos de tumores asociados con alelos HLA-I frecuentes en la población del Valle Central de Costa Rica

Diego Morazán-Fernández¹, José Arturo Molina-Mora²

Morazán-Fernández, D.; Molina-Mora, J.A. Colorectal cancer vaccines: in silico identification of tumor-specific antigens associated with frequent HLA-I alleles in the costarican Central Valley population *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 83-92.

 <https://doi.org/10.18845/tm.v35i8.6458>

1 Laboratorio clínico, Hospital Nacional de Niños, San José. Costa Rica
2 Facultad de Microbiología, Universidad de Costa Rica, San José. Costa Rica.
Centro de Investigación de Enfermedades Tropicales and Facultad de Microbiología, Universidad de Costa Rica, San José. Costa Rica.
E-mail: jose.molinamora@ucr.ac.cr
 <https://orcid.org/0000-0001-9764-4192>

Keywords

Neopeptide; Colorectal cancer; Cancer vaccine; Single nucleotide variants; HLA.

Abstract

Colorectal cancer is a complex disease in which uncontrolled growth of abnormal cells occurs in the large intestine (colon or rectum). The study of tumor-specific antigens (neoantigens), molecules that interact with the immune system, has been extensively explored as a possible therapy called in silico cancer vaccine. Cancer vaccine studies have been triggered by the current high-throughput DNA sequencing technologies. However, there is no universal bioinformatic protocol to study tumor-antigens with DNA sequencing data.

We propose a bioinformatic protocol to detect tumor-specific antigens associated with single nucleotide variants (SNVs) or “mutations” in colorectal cancer and their interaction with frequent HLA alleles (complex that present antigens to immune cells) in the Costa Rican Central Valley population. We used public data of human exome (DNA regions that produce functional products, including proteins). A variant calling analysis was implemented to detect tumor-specific SNVs, in comparison to healthy tissue. We then predicted and analyzed the peptides (protein fragments, the tumor specific antigens) derived from these variants, in the context of its affinity with frequent alleles of HLA type I of the Costa Rican population.

We found 28 non-silent SNVs, present in 26 genes. The protocol yielded 23 strong binders peptides derived from the SNVs for frequent alleles (greater than 8%) for the Costa Rican population at the HLA-A, B and C loci. It is concluded that the standardized protocol was able to identify neoantigens and this can be considered a first step for the eventual design of a colorectal cancer vaccine for Costa Rican patients. To our knowledge, this is the first study of an in silico cancer vaccine using DNA sequencing data in the context of the Costa Rican HLA alleles.

Palabras clave

Neopéptido; cáncer colorrectal; vacuna contra el cáncer; variantes de un solo nucleótido; HLA.

Resumen

El cáncer colorrectal es una enfermedad compleja en la que se produce un crecimiento descontrolado de células anormales en el intestino grueso (colon o recto). El estudio de antígenos específicos de tumor (neoantígenos), moléculas que interactúan con el sistema inmunológico, se ha explorado ampliamente como una posible terapia llamada vacuna contra el cáncer en silico. Los estudios de vacunas contra el cáncer han sido impulsados por las tecnologías actuales de secuenciación de ADN de alto rendimiento. Sin embargo, no existe un protocolo bioinformático universal para estudiar antígenos tumorales con datos de secuenciación de ADN.

Proponemos un protocolo bioinformático para detectar antígenos específicos de tumores asociados con variantes de un solo nucleótido (SNV) o “mutaciones” en el cáncer colorrectal y su interacción con alelos HLA frecuentes (complejo que presenta antígenos a las células inmunes) en el Valle Central de Costa Rica. población. Usamos datos públicos del exoma humano (regiones de ADN que producen productos funcionales, incluidas proteínas). Se implementó un análisis de llamada variante para detectar SNV específicos del tumor, en comparación con el tejido sano. A continuación, predecimos y analizamos los péptidos (fragmentos de proteínas, los antígenos específicos del tumor) derivados de estas variantes, en el contexto de su afinidad con los alelos frecuentes de HLA tipo I de la población costarricense.

Encontramos 28 SNV no silenciosos, presentes en 26 genes. El protocolo produjo 23 péptidos ligantes fuertes derivados de los SNV para alelos frecuentes (más del 8%) para la población costarricense en los loci HLA-A, B y C. Se concluye que el protocolo estandarizado logró identificar neoantígenos y esto puede considerarse un primer paso para el eventual diseño de una vacuna contra el cáncer colorrectal para pacientes costarricenses. Hasta donde sabemos, este es el primer estudio de una vacuna contra el cáncer en silico que utiliza datos de secuenciación de ADN en el contexto de los alelos HLA de Costa Rica.

Introduction

Colorectal cancer is a complex disease in which uncontrolled growth of ab-normal cells occurs in the large intestine (colon or rectum). In 2018, colorectal cancer worldwide reached an incidence of 1 849 518 cases and 880 792 deaths, occupying the third and second cause of cancer incidence and deaths respectively [1]. Cancer vaccines and others immunotherapies have emerged as an alternative to treat this type of cancer. As summarized in Fig. 1, there is evidence that immune system cells, such as CD8+ cytotoxic lymphocytes, can recognize tumor-specific peptides or proteins fragments derived from gene mutations or variants in malignant cells. In the immune response, these neopeptides are presented by cells (cancer or normal) to CD8+ cells. The antigen presentation occurs with the participation of HLA, the human major histocompatibility complex in tumor (and normal) cells, and the TCR (T-cell receptor) in CD8+ cells. If the antigen is tumor-specific, the death of tumor cells is induced [2].

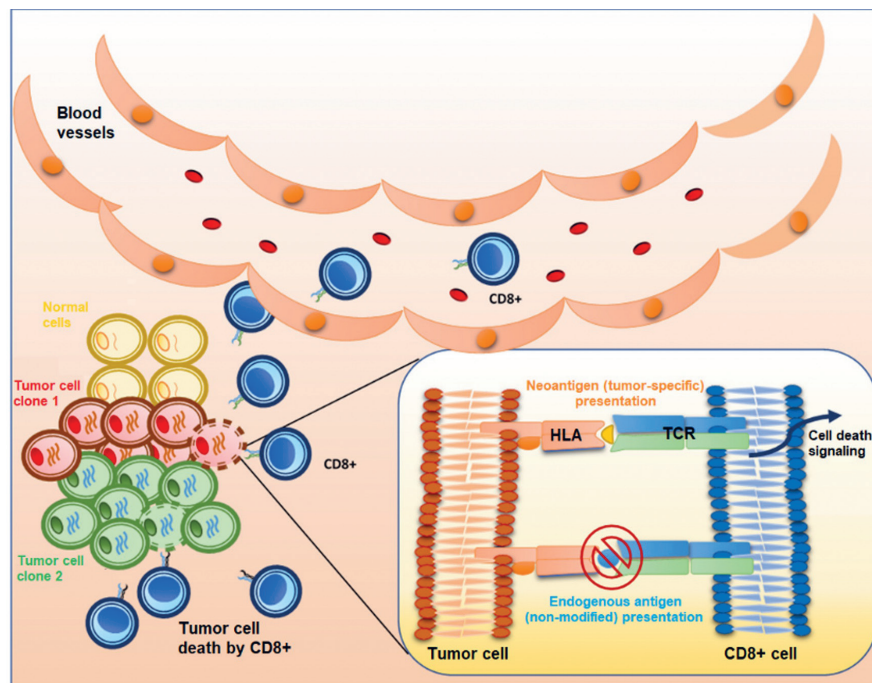


Figure 1. Conceptualization of the biological meaning of the neoantigens identification in cancer vaccines. The tumor-specific antigens (neoantigens) are originated from variants (mutations) in the genes of tumor cells, which are selected as candidate peptides for the vaccine. Peptides that are predicted to be presented in the HLA to cytotoxic CD8+ cells (HLA-antigen-TCR), are kept for the subsequent analyses. If the antigen is an endogenous antigen (non-modified and also present in normal cells), it is excluded. The eventual signaling in the immune response will induce death of tumor cells. Own elaboration.

With the advance of high-throughput DNA sequencing technologies, cancer vaccines and tumor/ immune system interactions cells have been extensively studied [3-5]. Cancer vaccine studies must consider HLA alleles (“gene versions”) that are frequent in the target population, the HLA type (type I or HLA-I for interaction with CD8+ cells), the peptide size (between 8-11 amino acids for HLA-I) and others [6]. However, the data complexity and the specific steps do not allow the application of universal protocols. Benchmark strategies are recommended for specific cases and populations [7-9]. Thus, the aim of this work was to implement a bioinformatic protocol of an in silico cancer vaccine to identify tumor-specific peptides in colorectal cancer considering the context of frequent HLA-I alleles from the Costa Rican Central Valley Population.

Methods

General workflow of the analysis is shown in Fig. 2.

Sequence data of tumor samples and pre-processing

Raw data from chromosome 1 of 7 exomes from human colorectal cancer (4 primary tumor and 3 from liver metastasis) and normal mucosa samples (all from same patient) were downloaded from NCBI Bioproject PRJNA271316 CHET9 experiment (sequenced by Illumina® technology, access: www.ncbi.nlm.nih.gov/bioproject/PRJNA271316). The exome reads were downloaded in Fastq format for subsequent analyzes. Quality control was done using FastQC, and trimming of reads was done with Trimmomatic, including filtering by quality and adapters removal [10].

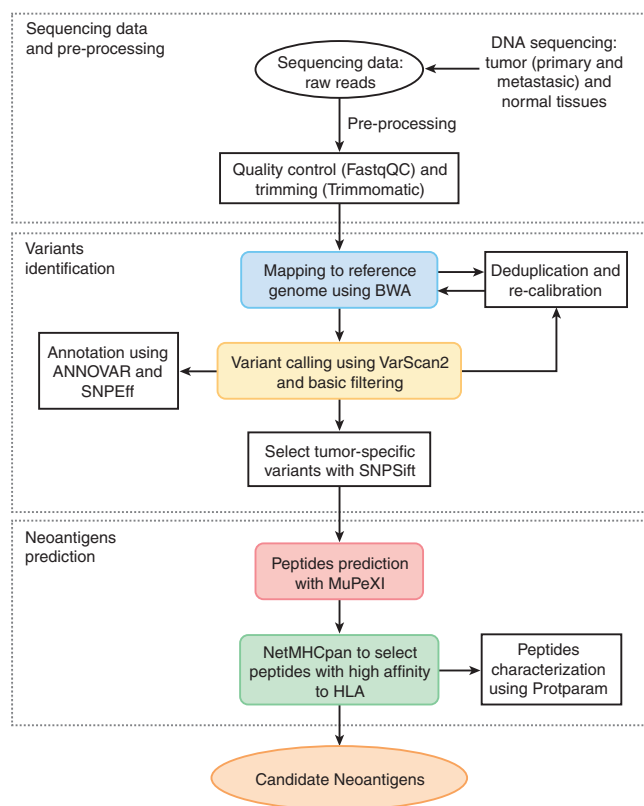


Figure 2. Bioinformatic pipeline to identify neoantigens in primary and metastatic tumor samples of colorectal cancer. The protocol includes three main steps: pre-processing, variant calling and neoantigens (peptides) identification.

Mapping to human genome and Variant Calling analysis

To map reads to the human reference genome (version hg19), the BWA program [11] was used with default parameters. Samtools [12] was used for extra data manipulation as usual. Picard (<http://broadinstitute.github.io/picard/>) was used to evaluate the quality of the mapping. VarScan2 program [13] was used to detect SNVs with quality Q higher than 20 in all samples; other parameters were used by default in the deduplication and re-calibration steps. In order to select mutations present in all the tumor samples, VCFlib (<https://github.com/vcflib/vcflib>) was used. Subsequently, the SnpEff program [14] was used to evaluate and compare the variants between samples. SnpSift [15] was used to filter all those SNVs that were in the normal mucosa sample to finally obtain the tumor exclusive variants. Finally, these mutations were annotated with ANNOVAR [16], obtaining the description of the SNVs and their possible effects on genes. The Atlas repository of colorectal cancer was used to see the involvement of the genes with selected SNVs and the possible metabolic or signaling pathways affected [17].

Putative tumor-specific peptides (neoantigens) identification

Selected tumor-specific variants were analyzed with MuPeXi package [18], to generate the peptides (8-11 amino acid residues) to the different alleles of HLA type I and the possible priorities of related mutant peptides. We chose those HLA type I alleles with frequencies greater than 8% for the Costa Rican Central Valley population [19], using Allele Frequency Net Database (<http://www.allelefrequencies.net/default.asp>). Due to HLA data had a resolution of two-digits, HLA supertypes (four-digits) were considered.

Mutant peptides were analyzed with the NetMHCpan program [20], to establish the affinity of these peptides to the respective alleles of HLA type I by means of the rank percentile (% rank). The 0.5% threshold was used for strong binders. Finally, the characterization of peptides was done using the Protparam tool of the ExPasy server [21], which provides physicochemical properties.

Results and discussion

Identification of neoantigens or tumor-specific antigens was described in the end of the last century [22]. DNA sequencing technologies and bioinformatic protocols have triggered the understanding of the interactions of tumor cells with the immune system, including the design of cancer vaccines. This strategy is not suitable for all cancer types due to a relatively high mutation rate is expected to generate new antigens. In colorectal cancer, melanoma and other tumor types, a high number of neoantigens is associated with patient response to immune therapies, making them suitable for cancer vaccines.

In this context, to detect possible variants with potential to induce tumor-specific peptides in samples of colorectal tumors, a bioinformatic protocol was implemented with samples of human colorectal cancer for chromosome 1. We identified of 395 SNVs shared by all tumor samples (and absent in the normal mucosa sample).

Table 1. Strong binding peptides for most frequent HLA alleles of the Costa Rican population. Candidates are derived from the selected SNVs in the primary and metastatic tumor samples (pI: Isoelectric point, GRAVY: Grand average of hidropathy).

Alleles	Peptides	pI	Acidic nature	GRAVY	Lenght (AA)
A*02:01	KLVGFSLDV	5.84	acidic	1.133	9
	LLIHCDAYL	5.80	acidic	1.356	9
	YLHSPMYFFI	6.74	neutral	0.76	10
	ILLIHCDAYL	5.08	acidic	1.67	10
A*03:01	AMNILEINEK	4.53	acidic	-0.14	10
	GSSFYALEK	6.00	neutral	-0.256	9
A*24:02	AYLHSPMYFF	6.78	neutral	0.49	10
	AYLHSPMYFFI	6.78	neutral	0.855	11
	CDAYLHSPMYF	5.08	acidic	0.1	11
	DAYLHSPMYF	5.08	acidic	-0.14	10
	DAYLHSPMYFF	5.08	acidic	0.127	11
	RWYSTPSSYL	8.59	basic	-0.89	10
	YLHSPMYFF	6.74	neutral	0.344	9
B*07:02	GPLSSEKAAM	6.00	neutral	-0.17	10
B*35:01	DAYLHSPMY	5.08	acidic	-0.467	9
B*40:01	AEINILEI	3.80	acidic	1.075	8
	EEKLVGFSL	4.53	acidic	0.278	9
	LEEKLVGFSL	4.53	acidic	0.63	10
	RELTHLREKL	8.75	basic	-1.24	10
	SEKAAMNIL	5.52	acidic	0.233	9
B*44:02	AEINILEI	3.80	acidic	1.075	8
C*03:02	KAAMNILEI	6.00	neutral	0.822	9
C*04:01	AYLHSPMYF	6.78	neutral	0.233	9
	RWYSTPSSYL	8.59	basic	-0.89	10
	WYSTPSSYL	5.52	acidic	-0.489	9
	YLHSPMYFF	6.74	neutral	0.344	9
C*06:02	WYSTPSSYL	5.52	acidic	-0.489	9
C*07:01	WYSTPSSYL	5.52	acidic	-0.489	9

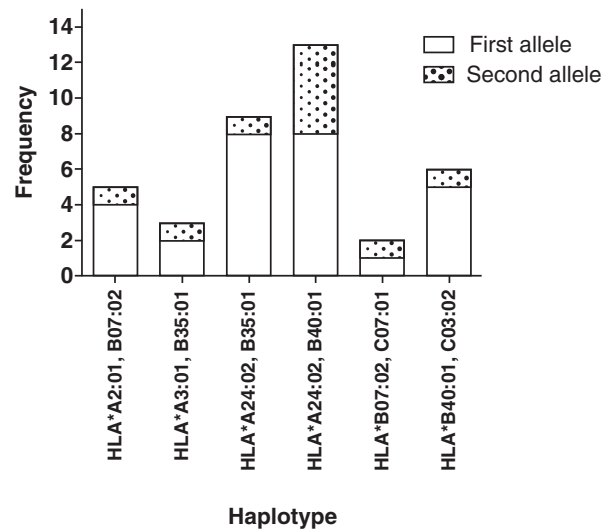


Figure 3. Number of strong binding peptides for different HLA I haplotypes from Costa Rican population. Peptides are derived from the selected SNVs in the primary and metastatic tumor samples of the study.

Despite this, only 28 non-silent SNPs in exonic regions were identified. They were selected to predict neoantigens. These variants were distributed in 26 genes, most of them related to cell cycle control. Interestingly, multiple of these genes, including NOTCH, OBSCN, HSPG2, and NBPF, have been demonstrated to be expressed in colorectal cancer [23-25]. Further analyses are required to assess the role of these specific genes in the development of the disease and the immune response, including studies of driving genes (important for the tumor survival), transcriptomic/proteomic profiling, epigenetic mechanisms or other genomic variants.

Regarding the SNVs identification, we used VarScan2 to call variants (mutations). This software has been shown to have a high sensitivity (81-100%) in the variant calling analysis, in particular with cancer data [26, 27]. In addition, the use of a normal tissue sample to filter variants has been advised by several studies to reduce false positives (such as germline mutations) [28, 29], as we did here. Confirmation strategies are required to validate identified variants (for example using Sanger technology to re-sequence genomic regions). Also, other approaches to call variants can be used to assess other DNA data, to deal with complexity, other sequencing technologies, and biological variability [30].

On the other hand, for HLA of the Costa Rican Central Valley population, information only had a two-digits resolution. Thus, we used the HLA supertypes (A*01, A*02, A*03, A*24 and B*07) or the representative allele of the group to predict interactions (four-digits resolution as required). The length of tumor-specific peptides was considered between 8-11 amino acids, which is the known peptide size that can be presented by HLA-I. It has been suggested that simultaneous testing of peptides between 8 and 11 amino acid residues is advisable in peptide prediction for putative cancer vaccines, due to the “core” or the common sequence (8 amino acids) may have greater versatility in different fragments [31]. This allowed to test several possible peptides of the same variant and to improve the searching for possible neoantigens.

Considering the above, 23 candidate peptides were found to have affinity against the most prevalent Costa Rican Central Valley HLA I alleles (Table 1). Some peptides were identified to be presented by different alleles of the HLA, such as AYLHSPMYF (two different loci HLA A*2402 and HLA * C0401), however, most peptides are predicted to be presented in only one allele.

Regarding the number of peptides and haplotypes, the HLA-A locus was the most significant, contributing to most of the high affinity peptides for the antigenic presentation (Fig. 3 and Table 1). For example, up to nine peptides can be presented in the case of haplotype A*2402/B*3501, in which the first allele is responsible for the presentation of nine peptides.

We also analyzed the physicochemical nature of the peptides (Table 1). Many of them are acidic peptides and, based on the hydrophobicity index, most of them have a slight hydrophobicity. Some “core” peptides are found in several peptides, such as AYLHSPMY which is present as part of seven candidate peptides of different HLA alleles (HLA-A*24:02, B*35:01 and C*04:01).

In the context of the selection of peptides, the algorithms use metrics such as IC50 value (concentration necessary to displace 50% of the classical peptide bound to an HLA allele) [32, 33]. However, it has been shown that it can yield a greater number of false positives in the detection of neopeptides, since there are some alleles that can bind a significant percentage of random peptides [20]. This motivated the present investigation to use only peptides with strong affinities (rank < 0.5%).

Other biologic factors should be considered to identify a possible neoantigens and not included in our study. For example, protein degradation is possible at the endoplasmic reticulum [34]. In addition, peptides affinity must be validated using *in vitro* assays with antigen-presenting cells of the Costa Rican population, for example the Enzyme-linked Immunospot Assay (ELISPOT) [35].

Conclusions

In summary, our approach was able to find 23 candidate neoantigens in all samples of colorectal cancer to eventually create a vaccine. The contribution of this study is the development of the first bioinformatic protocol to detect tumor-specific antigens in the context of the HLA alleles of the Costa Rican population. We hope that this study serve as a basis for future studies with therapeutic uses for cancer vaccines against colorectal tumors or other cancer types.

References

- [1] World Health Organization, Colorectal Cancer Statistics: 2020. Available from: https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf
- [2] Palucka AK, Coussens LM. The Basis of Oncoimmunology. *Cell* [Internet]. 2016;164(6):1233–47. Available from: <http://dx.doi.org/10.1016/j.cell.2016.01.049>
- [3] Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet*. 2016;17:441–58.
- [4] Spencer DH, Zhang B, Pfeifer J. Single Nucleotide Variant Detection Using Next Generation Sequencing. In: Kulkarni, S.; Pfeifer J, editor. *Clinical Genomics* [Internet]. Elsevier Inc.; 2014. p. 109–27. Available from: <http://dx.doi.org/10.1016/B978-0-12-404748-8.00008-3>
- [5] Ding L, Wendl MC, McMichael JF, Raphael BJ. Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet* [Internet]. 2014;15(8):556–70. Available from: <http://dx.doi.org/10.1038/nrg3767>
- [6] Liu XS, Mardis ER. Applications of Immunogenomics to Cancer. *Cell*. 2017;168(4):600–12.
- [7] Su Z, Ning B, Fang H, Hong H, Perkins R, Shi L. Next-generation sequencing and its application in molecular diagnostics. *Expert Rev Mol Diagn*. 2011;11(3):1–16.
- [8] Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics* [Internet]. 2016;107(1):1–8. Available from: <http://dx.doi.org/10.1016/j.ygeno.2015.11.003>
- [9] Rizzo JM, Buck MJ. Key principles and clinical applications of “next-generation” DNA sequencing. *Cancer Prev Res (Phila)* [Internet]. 2012 Jul 1 [cited 2018 Apr 25];5(7):887–900. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22617168>
- [10] Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, et al. Manipulation of FASTQ data with galaxy. *Bioinformatics*. 2010;26(14):1783–5.

- [11] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(15):1754–60.
- [12] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
- [13] Koboldt DC, Zhang Q, Larson DE, Shen D, Mclellan MD, Lin L, et al. VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22:568–76.
- [14] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*. 2012;6(2):80–92.
- [15] Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program , SnpSift. *Front Genet*. 2012;3(315):1–9.
- [16] Wang K, Li M, Hakonarson H. ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):1–7.
- [17] Chisanga D, Keerthikumar S, Pathan M, Ariyaratne D, Kalra H, Boukouris S, et al. Colorectal cancer atlas : An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues. *Nucleic Acids Res*. 2016;44(D1):969–74.
- [18] Mette A, Morten B, Hadrup SR. MuPeXI : prediction of neo - epitopes from tumor sequencing data. *Cancer Immunol Immunother*. 2017;66(9):1123–30.
- [19] Arrieta-Bolaños E, Maldonado-Torres H, Dimitriu O, Hoddinott MA, Fowles F, Shah A, et al. HLA-A, -B, -C, -DQB1, and -DRB1,3,4,5 allele and haplotype frequencies in the Costa Rica Central Valley Population and its relationship to worldwide populations. *Hum Immunol*. 2011;72(1):80–6.
- [20] Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med [Internet]*. 2016;8(1):1–9. Available from: <http://dx.doi.org/10.1186/s13073-016-0288-x>
- [21] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: *Wlaker JM, editor. The Proteomics Protocols Handbook*. Human Press; 2005. p. 571–608.
- [22] Emens LA, Ascierto PA, Darcy PK, Demaria S, Eggermont AMM, Redmond WL, et al. Cancer immunotherapy: Opportunities and challenges in the rapidly evolving clinical landscape [Internet]. Vol. 81, *European Journal of Cancer*. 2017 [cited 2018 May 4]. p. 116–29. Available from: https://ac.els-cdn.com/S0959804917309188/1-s2.0-S0959804917309188-main.pdf?_tid=852cff99-60b9-4e00-9a8c-eca592b4c0be&acdnat=1525459824_17fe72594087a4b1c451941055acd504
- [23] Andries V, Vandepoele K, Staes K, Bex G, Bogaert P, Isterdael G, et al. NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. *BMC Cancer*. 2015;15(1):1–25.
- [24] Shriver M, Stroka KM, Vitolo MI, Martin S, DL Huso DL, Konstantopoulos K, et al. Loss of giant obscurins from breast epithelium promotes epithelial-to-mesenchymal transition, tumorigenicity and metastasis. *Oncogene*. 2015;34(32):4248–59.
- [25] Liao W, Li G, You Y, Wan H, Wu Q, Wang C, et al. Antitumor activity of Notch-1 inhibition in human colorectal carcinoma cells. *Oncol Rep*. 2018;39(3):1063–71.
- [26] Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, et al. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. *J Mol Diagnostics [Internet]*. 2014;16(1):75–88. Available from: <http://dx.doi.org/10.1016/j.jmoldx.2013.09.003>
- [27] Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J [Internet]*. 2018;16:15–24. Available from: <https://doi.org/10.1016/j.csbj.2018.01.003>
- [28] Bae JM, Kim JH, Kang GH. Molecular subtypes of colorectal cancer and their clinicopathologic features, with an emphasis on the serrated neoplasia pathway. *Arch Pathol Lab Med*. 2016;140(5):406–12.
- [29] Wang Y, Liping GUO, Feng L, Zhang W, Xiao T, Xuebing DI, et al. Single nucleotide variant profiles of viable single circulating tumour cells reveal CTC behaviours in breast cancer. *Oncol Rep*. 2018;39(5):2147–59.
- [30] Liu ZK, Shang YK, Chen ZN, Bian H. A three-caller pipeline for variant analysis of cancer whole-exome sequencing data. *Mol Med Rep*. 2017;15(5):2489–94.
- [31] Luo H, Ye H, Ng HW, Shi L, Tong W, Mattes W, et al. Understanding and predicting binding between human leukocyte antigens (HLAs) and peptides by network analysis. *BMC Bioinformatics [Internet]*. 2015 [cited 2018 May 16];16(13). Available from: <http://www.biomedcentral.com/1471-2105/16/S13/S9>

- [32] Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep.* 2016;15(4):857–65.
- [33] Karasaki T, Nagayama K, Kuwano H, Nitadori J ichi, Sato M, Anraku M, et al. An Immunogram for the Cancer-Immunity Cycle: Towards Personalized Immunotherapy of Lung Cancer. *J Thorac Oncol* [Internet]. 2017;12(5):791–803. Available from: <http://dx.doi.org/10.1016/j.jtho.2017.01.005>
- [34] Fleri W, Paul S, Dhanda SK, Mahajan S, Xu X, Peters B, et al. The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front Immunol.* 2017;8(278):1–16.
- [35] Kato T, Matsuda T, Ikeda Y, Park J-H, Leisegang M, Yoshimura S, et al. Effective screening of T cells recognizing neoantigens and construction of T-cell receptor-engineered T cells. *Oncotarget.* 2018;9(13):11009–19.





Automatic diagnosis of lower back pain using gait patterns

Diagnóstico automático del dolor lumbar mediante patrones de marcha

Chandrasen Pandey¹, Neeraj Baghel²,
Malay Kishore-Dutta³, Carlos M. Travieso⁴

Pandey, CH.; Baghel. N.; Kishore-Dutta M.; Travieso, C. Automatic diagnosis of lower back pain using gait patterns. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 93-100.

 <https://doi.org/10.18845/tm.v35i8.6459>

- 1 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University. India. E-mail: developer.chandrasen@gmail.com
 <https://orcid.org/0000-0002-7031-1619>
- 2 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University. India. E-mail: nbaghel777@gmail.com
 <https://orcid.org/0000-0002-0081-6224>
- 3 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University. India. Correo electrónico: malaykishoredutta@gmail.com
 <https://orcid.org/0000-0003-2462-737X>
- 4 Signals and Communications Department, IDeTIC, University of Las Palmas de Gran Canaria, Las Palmas. Spain. E-mail: carlos.travieso@ulpgc.es
 <https://orcid.org/0000-0002-4621-2768>

Keywords

Gait Analysis; Back Pain; Support vector machine.

Abstract

Back pain is a common pain that mostly affects people of all ages and results in different types of disorders such as Obesity, Slipped disc, Scoliosis, and Osteoporosis, etc. The diagnosis of back pain disorder is difficult due to the extent affected by the disorder and exact biomechanical factors. This work presents a machine learning method to diagnose these disorders using the Gait monitoring system. It involves support vector machines that classify between lower back pain and normal, on the bases of 3 Gait patterns that are integrated pressure, the direction of progression, and CISP-ML. The proposed method uses 13 different features such as mean and standard deviation, etc. recorded from 62 subjects (30 normal and 32 with lower back pain). The features alone resulted in higher leave-one-out classification accuracy (LOOCV) 92%. The proposed method can be used for automatically diagnosing the lower back pain and its gait effects on the person. This model can be ported to small computing devices for self-diagnosis of lower back pain in a remote area.

Palabras clave

Análisis de la marcha; dolor de espalda; máquina de vectores de apoyo.

Resumen

El dolor de espalda es un dolor común que afecta principalmente a personas de todas las edades y da como resultado diferentes tipos de trastornos como obesidad, deslizamiento de disco, escoliosis y osteoporosis, etc. El diagnóstico del trastorno de dolor de espalda es difícil debido a la extensión del trastorno y factores biomecánicos exactos. Este trabajo presenta un método de aprendizaje automático para diagnosticar estos trastornos mediante el sistema de monitorización de la marcha. Se trata de máquinas de vectores de apoyo que clasifican entre lumbalgia y normal, sobre la base de 3 patrones de marcha que son la presión integrada, la dirección de progresión y CISP-ML. El método propuesto utiliza 13 características diferentes, como la desviación media y estándar, etc. registrado de 62 sujetos (30 normales y 32 con dolor lumbar). Las características por sí solas dieron como resultado una mayor precisión de clasificación de dejar uno fuera (LOOCV) del 92%. El método propuesto se puede utilizar para diagnosticar automáticamente el dolor lumbar y sus efectos sobre la marcha en la persona. Este modelo se puede transferir a pequeños dispositivos informáticos para el autodiagnóstico del dolor lumbar en un área remota.

Introduction

Back pain, commonly known as backache, is pain occurs mostly in the lower back section. The back has different parts that are classified into neck pain, also known as cervical, middle back pain also known as thoracic, lower back pain is also known as lumbar or coccydynia (tailbone or sacral pain) are part of the pain. [7, 9]. At the time of occurrence, Back pain may be chronic or acute, sub-acute, depending on the part. The pain may be of different types such as a dull ache piercing pain or shooting, or sometimes it gives a burning sensation. These types of pain can be analyzed with the help of gait data, with the help of gait analysis take at the events as one continuous gait motion [3]. After all, 15% of the average person's total weight was situated in the hip, thigh, lower leg, and foot comprise around. So, to understand how this weight is

transferred during walking is critical to preventing RSI and other injuries [7]. The study of Podiatric biomechanics uses the historic data of gait analysis. Musculoskeletal problems can occur with Gait abnormalities, such as back discomfort and changes in posture [3]. According to APA Frymoyer [9], Sports person and cross-country skiers are the Patients with moderate low-back pain has been more often pretentious when compared with the asymptomatic men with severe low back pain. Otherwise, there is no variability related to sports activity. The previous method use MRI, X-ray, and CT-Scan [10], which is costly and time taking process. The main object of this paper is to develop an easy and low-cost, lower back pain detection system using machine learning.

In this paper, we proposed an automatic diagnostic tool for a lower back pain using gait patterns based on the machine learning classifier. We have explored the features like a combination of integrated pressure, the centre of pressure, and the degree of a proposition for the lower back pain. We have also explored the statistical feature of these gait patterns to diagnose lower back pain using the SVM.

This paper is organized as follows. Section 2 describes the data uses and equipment details. Section 3 consists of a proposed methodology with data processing and feature extraction. In Section 4 described the experimental results and Section 5 describes the conclusion.

Dataset used

The Gait dataset is collected from Dr. Shakuntala Misra National Rehabilitation University, India, and was found 30-40% of people are infected with a different kind of back pain. The dataset is collected from The Proto Kinetics Zeno Walkway Gait Analysis System. Also, Proto Kinetics Movement Analysis Software is used that can effortlessly record and outputs temporal, spatial, and pressure measurements for various protocols. All data are stored in the form of an XML format that can be used to examine the gait of a subject.

The Gait monitor system contains a pressure sensing pad with dual control the sensor is embedded in a base layer, surrounded by a bevelled edge, the sensing pad connected with a system that analyses and record the sensor values and a camera set that is used to



Figure 1. A person of Age 42 diagnosed with back pain walking on the Gait monitor sensing pad and all required gait values are stored in the system with motion recording with the camera.

Table 1. Gait Dataset Explanation

Gait Values	Gender	Total subjects
Normal	Male -12, Female-18	30
Lower Back Pain	Male -16, Female-16	32
Total	Male-28, Female-34	62

The dataset was divided into two different categories, such as, 'Normal' and 'Lower Back pain'. The data consists of a total of 62 subjects (28 males and 34 females) where 30 subjects (12 Males and 18 Females) with Normal gait values and 32 subjects (16 Males and 16 females) suffering from back pain as shown in table 1.

Methodology

In this section, the proposed machine learning method for the identification of lower back pain using the gait pattern is described. It involves the processing of gait data of normal subjects and lowers back pain subject. To extract useful information from it for the training of a machine-learning algorithm to decide whether the gait data is lower back pain or not. It involves the steps of data processing, feature extraction, class labelling, etc., as discussed next.

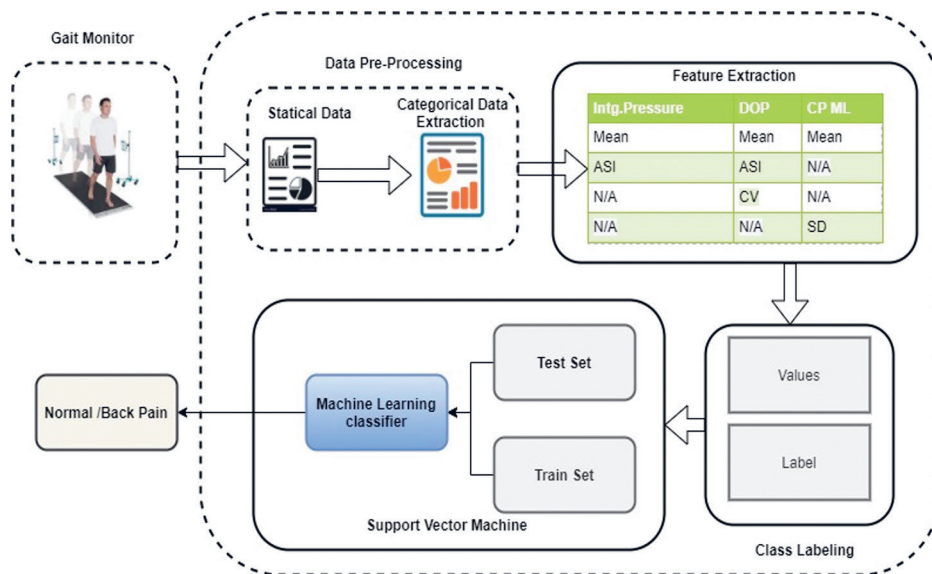


Figure 2. Flow diagram of a proposed algorithm.

Figure 2 gives the process flow obeyed in the proposed machine learning process for the diagnosis of lower back pain. Data generated from the gait monitor system is stoical data. It is preprocessed to extract categorical data from it, which discriminate values between lower back pain data and normal data. From the experimental study, three different gait patterns have been analyzed (i) the integrated pressure, (ii) the direction of progress, and (iii) the centre of propagation. Further, the statistical features of these categories have been used for training and testing of the proposed machine learning algorithm.

Feature extraction

The feature extraction is used to select the discriminate values that can be used to classify the normal gait data with lower back pain data. The proposed method has taken 3 gait patterns (Integrated pressure, the direction of progress, a centre of pressure) for the analysis. Integrated pressure (IP) measurement quantitatively characterizes the contact between the ground and foot. Clinical judgment of foot function mostly relies on the analysis of gait pressure and pressure-related framework under specific regions of the foot [12]. The shoulder joints, hip joints and the spine are known as “centre of mass” (COM). The median (in each coordinate) of these variables is considered the contemporaneous direction of progress (DOP) [13]. CISP-ML describes the shift of the Center of Pressure (CP) from the anterior-posterior axis of the Center of Origin Deviation is picked with a negative sign (-) if the positive (+) for an anterior shift and CISP-AP is shifted posteriorly. If it has deviated to the left a positive (+) value is applied and CISP-ML is shifted to the right it is picked with a negative sign (-) [14]. The statistical feature of these 3 gait patterns such as mean, left foot means, right foot means, accumulative swing index (ASI), coefficient of variation (CV), standard deviation (SD), SD left foot (SD LEFT), SD right foot (SD RIGHT), has been used for further analysis.

The proposed method has taken mean and variance from each sample data to determine the p values of the discriminant feature. The Extracted features of sample data and its corresponding p values are explained in table 2.

Table 2. Extracted features of sample data and its corresponding p-values.

S.no	Feature	Lower Back Pain Sample Value (Mean±Variance)	Normal Sample Value (Mean±Variance)	P values
1	IP Mean	179.4 ± 1822.1212	125.6086 ± 126.9452	2.33E-07
2	IP Mean(L)	178.9199 ± 1486.0614	165.8322 ± 1571.6691	0.04755059
3	IP Mean(R)	186.0457 ± 2443.3185	157.6840 ± 937.9480	0.001705977
4	IP ASI	-3.8332 ± 11.8333	4.5858 ± 6.61844	0.002133478
5	DoP Mean	-9.5851 ± 544.3209	-6.8516 ± 1980.9582	0.319169781
6	DoP Mean(L)	-28.8701 ± 505.89008	0.16193 ± 3667.4503	0.758613854
7	DoP Mean(R)	21.9554±1701.9548	-14.25426±1837.7280	0.571749924
8	Dop ASI	38.7139±3199.2950	-30.8834±9224.8274	0.939697302
9	Dop CV	-3796.234±278598765.71	15.7230±155342.5139	0.417152495
10	CPIML Mean	-2.1354±36.0454	-1.26973±5.02719	0.032416617
11	CPIML SD	-174.7287±201200.4723	845.0284±2693582.9156	0.15324072
12	CPIML SD(L)	-548.0343±2112559.7763	199.4138±633361.2656	0.353923927
13	CPIML SD(R)	-158.0511±99472.8081	-118.4832±510053.1774	0.789085173

In table 3 shows the features of the sample to find the most suitable features to implement. The mean and variance of these features are calculated for both normal and lower back pain suffering subjects. The P-values are calculated to check the prominent features using the probability of features.

Machine learning algorithm

A powerful supervised learning method is Support vector machine (SVM). Kernel function that is used in SVM plays an important for proper training of the SVM. There are various application in 1D signal [15,16]. In this work Linear Kernel is used for binary classification. The performance of classifiers is measured with different parameters like Sensitivity (Recall), Precision, Specificity, F1-Score, False Positive Rate, and Accuracy are used which are calculated as follows:

$$\text{Sensitivity(Recall)} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{F1 Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (4)$$

Here true positive is denoted as TP, the total number of predicted values from a class A. True Negative is denoted by TN, which means the total number of predicted samples from class B. False Positive denoted by FP which means to express the total number class A gait values which are classified as class B. False-negative denoted by FN that means to defined the total number of class B Gait samples which are classified as class A.

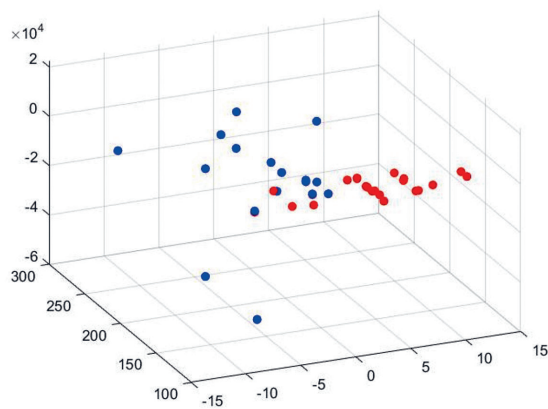


Figure 3. 3D Scatter plot of two Classes

Experimental Results

The proposed method used a Gait dataset with a total of 21 Statistical values for experimentation. It is classified into two different categories as healthy and lower back pain. The number of samples from healthy and lower back pain categories was 30 and 32, respectively. Due to a small dataset the “Leave-one-out-cross-validation (LOOCV)” has been used for cross-validation. The proposed method has been trained on SVM, knn, naïve Bais. The confusion matrix of these classifiers is shown in table 3.

Table 3. Testing set classification confusion matrix

		KNN classifier		Naïve bias classifier		SVM classifier	
		Normal	Lower Back Pain	Normal	Lower Back Pain	Normal	Lower Back Pain
True Label	Normal	32	0	26	6	30	2
	Lower Back Pain	30	0	2	28	3	27
		Predicted values		Predicted values		Predicted values	

In table 3 shows the confusion matrix of different classifiers (a) Knn classify 32 normal and 0 lower back pain, (b) naïve bias classify 26 normal and 28 lower back pain and (c) SVM classifies 30 normal and 27 lower back pain correctly. The performance of each classifier is compared with various parameters and it's described in table 4.

Table 4. Comparative performance of different classifier

Class	TP	FP	FN	TN	Precision	Recall	F1 score	Accuracy
Performance of KNN classifier								
Normal	32	0	30	0	0.00	0.00	0.00	
Lower Back Pain					0.52	1.00	0.68	0.52
Average					0.26	0.5	0.34	
Performance of Naïve bias classifier								
Normal	26	6	2	28	0.82	0.93	0.87	
Lower Back Pain					0.93	0.81	0.87	0.87
Average					0.875	0.87	0.87	
Performance of SVM classifier								
Normal	30	2	3	27	0.93	0.90	0.92	
Lower Back Pain					0.91	0.94	0.92	0.92
Average					0.92	0.92	0.92	

Table 5. Comparison of the proposed method against existing method

Reference	Methodology Used	Dataset	Accuracy
oameng Ung[7]	Structural MRI Data Detects Chronic Low Back Pain	MRI Dataset	76%
K Ritwik[8]	Analysis of lower back pain disorder using deep learning	LBP x-ray data	65%
Proposed Method	Special domain statistical and textural features classification using a Binary class support vector machine.	62 Gait Dataset	92%

The comparison of our proposed work with a similar kind of work is shown in table 4. The oameng [7] used an MRI dataset to detect Low Back Pain with 76% accuracy. K Ritwik [8] used X-ray to detect lower back pain using Deep learning.

Conclusions

A classification technique was used to recognize lower back pain with gait patterns. The overall test result indicated that the proposed method using the SVM classifier was able to effectively diagnose lower back pain conditions from a small contribution of 92% accuracy. This was encouraging for the future application of SVMs in gait diagnostics as well as in the assessment of treatment arbitration, especially in the lower back pain inhabitant. Future work may include multi-classification of various types of pain and the use of deep learning method to improve accuracy for the diagnosis of pain.

References

- [1] A. J. Aljaaf, A. J. Hussain, P. Fergus, A. Przybyla and G. J. Barton, "Evaluation of machine learning methods to predict knee loading from the movement of body segments," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 5168-5173, DOI: 10.1109/IJCNN.2016.7727882.
- [2] D. T. H. Lai, P. Levinger, R. K. Begg, W. L. Gilleard and M. Palaniswami, "Automatic Recognition of Gait Patterns Exhibiting Patellofemoral Pain Syndrome Using a Support Vector Machine Approach," in IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 5, pp. 810-817, Sept. 2009, DOI: 10.1109/TITB.2009.2022927.
- [3] Pogorelc, B., Bosnić, Z. & Gams, M. Automatic recognition of gait-related health problems in the elderly using machine learning. *Multimed Tools Appl* 58, 333–354 (2012).
- [4] Pier Nicola Sergi, Winnie Jensen, Silvestro Micera, Ken Yoshida, In vivo interactions between tungsten micro-needles and peripheral nerves, *Medical Engineering & Physics*, Volume 34, Issue 6, 2012, Pages 747-755, ISSN 1350-4533.
- [5] R. Norman, R. Wells, P. Neumann, J. Frank, H. Shannon, M. Kerr, A comparison of peak vs cumulative physical work exposure risk factors for the reporting of low back pain in the automotive industry, *Clinical Biomechanics*, Volume 13, Issue 8, 1998, Pages 561-573, ISSN 0268-0033.
- [6] K. Kong and M. Tomizuka, "A Gait Monitoring System Based on Air Pressure Sensors Embedded in a Shoe," in IEEE/ASME Transactions on Mechatronics, vol. 14, no. 3, pp. 358-370, June 2009, DOI: 10.1109/TMECH.2008.2008803.
- [7] oameng Ung, Justin E. Brown, Kevin A. Johnson, Jarred Younger, Julia Hush, Sean Mackey, Multivariate Classification of Structural MRI Data Detects Chronic Low Back Pain, *Cerebral Cortex*, Volume 24, Issue 4, April 2014, Pages 1037–1044, <https://doi.org/10.1093/cercor/bhs378>
- [8] Kulkarni, K & Gaonkar, Abhijitsingh & Vijayarajan, Vijayan & Manikandan, K. (2017). Analysis of lower back pain disorder using deep learning. *IOP Conference Series: Materials Science and Engineering*. 263. 042086. 10.1088/1757-899X/263/4/042086.
- [9] APA Frymoyer, J W†; Pope, M H†; Clements, J H†; Wilder, D G†; MacPherson, B†; Ashikaga, T† Risk factors in low-back pain. An epidemiological survey., *JBJS*: Feb 1983 - Volume 65 - Issue 2 - p 213-218
- [10] <https://www.everydayhealth.com/back-pain/back-pain-diagnosis.aspx>, Last access on 21-07-2020
- [11] R.W. Soames, Foot pressure patterns during gait, *Journal of Biomedical Engineering*, Volume 7, Issue 2, 1985, Pages 120-126, ISSN 0141-5425, [https://doi.org/10.1016/0141-5425\(85\)90040-8](https://doi.org/10.1016/0141-5425(85)90040-8).
- [12] Giacomozzi C., Caravaggi P., Stebbins J.A., Leardini A. (2016) Integration of Foot Pressure and Foot Kinematics Measurements for Medical Applications. In: Müller B. et al. (eds) *Handbook of Human Motion*. Springer, Cham https://doi.org/10.1007/978-3-319-30808-1_186-1
- [13] M. Gabel, R. Gilad-Bachrach, E. Renshaw and A. Schuster, "Full body gait analysis with Kinect," 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, 2012, pp. 1964-1967, DOI: 10.1109/EMBC.2012.6346340.
- [14] <https://padulainstitute.com/education/articles/risk-fall-rof-intervention-affecting-visual-ego-center-gait-analysis-yoked-prisms/>, Last access on 21-07-2020
- [15] Baghel, N., Singh, D., Dutta, M. K., Burget, R., & Myska, V. (2020, July). Truth Identification from EEG Signal by using Convolution neural network: Lie Detection. In 2020 43rd International Conference on Telecommunications and Signal Processing (TSP) (pp. 550-553). IEEE.
- [16] Anjali Yadav, Anushikha Singh, Malay Kishore Dutta, and Carlos M. Travieso. "Machine learning-based classification of cardiac diseases from PCG recorded heart sounds." *Neural Computing and Applications*, pages 1-14. Springer, 2019.

Artificial Intelligence based Multi-sensor COVID-19 Screening Framework


Inteligencia artificial para el marco de detección COVID-19 multisensorial

Rakesh Chandra-Joshi¹, Malay Kishore-Dutta², Carlos M. Travieso³

Chandra-Joshi, R.; Kishore-Dutta, M.; Travieso, C.M. Artificial intelligence based multi-sensor covid-19 screening framework. *Tecnología en Marcha*. Tecnología en marcha. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 101-109.

 <https://doi.org/10.18845/tm.v35i8.6460>

1 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University, India. E-mail: rakeshchandraindia@gmail.com

 <https://orcid.org/0000-0003-1264-9010>

2 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University, India. E-mail: malaykishoredutta@gmail.com

 <https://orcid.org/0000-0003-2462-737X>

3 Signals and Communications Department, IDeTIC, University of Las Palmas de Gran Canaria, Las Palmas. Spain. E-mail: carlos.travieso@ulpgc.es

 <https://orcid.org/0000-0002-4621-2768>

Keywords

Convolutional Neural Network; COVID-19 detection; Deep Learning; Multi-sensor.

Abstract

Many countries are struggling for COVID-19 screening resources which arises the need for automatic and low-cost diagnosis systems which can help to diagnose and a large number of tests can be conducted rapidly. Instead of relying on one single method, artificial intelligence and multiple sensors based approaches can be used to decide the prediction of the health condition of the patient. Temperature, oxygen saturation level, chest X-ray and cough sound can be analyzed for the rapid screening. The multi-sensor approach is more reliable and a person can be analyzed in multiple feature dimensions. Deep learning models can be trained with multiple chest x-ray images belonging to different categories to different health conditions i.e. healthy, COVID-19 positive, pneumonia, tuberculosis, etc. The deep learning model will extract the features from the input images and based on that test images will be classified into different categories. Similarly, cough sound and short talk can be trained on a convolutional neural network and after proper training, input voice samples can be differentiated into different categories. Artificial based approaches can help to develop a system to work efficiently at a low cost.

Palabras clave

Red neuronal convolucional; detección COVID-19; aprendizaje profundo; sensor múltiple.

Resumen

Muchos países están luchando por los recursos de detección de COVID-19, lo que plantea la necesidad de sistemas de diagnóstico automáticos y de bajo costo que puedan ayudar a diagnosticar y que se pueda realizar una gran cantidad de pruebas rápidamente. En lugar de depender de un solo método, se pueden utilizar la inteligencia artificial y enfoques basados en múltiples sensores para decidir la predicción del estado de salud del paciente. La temperatura, el nivel de saturación de oxígeno, la radiografía de tórax y el sonido de la tos se pueden analizar para la detección rápida. El enfoque de múltiples sensores es más confiable y una persona puede ser analizada en múltiples dimensiones de características. Los modelos de aprendizaje profundo se pueden entrenar con múltiples imágenes de rayos X de tórax que pertenecen a diferentes categorías para diferentes condiciones de salud, es decir, saludable, COVID-19 positivo, neumonía, tuberculosis, etc. El modelo de aprendizaje profundo extraerá las características de las imágenes de entrada y en base a eso, las imágenes de prueba se clasificarán en diferentes categorías. De manera similar, el sonido de la tos y la conversación corta se pueden entrenar en una red neuronal convolucional y, después de un entrenamiento adecuado, las muestras de voz de entrada se pueden diferenciar en diferentes categorías. Los enfoques basados en materiales artificiales pueden ayudar a desarrollar un sistema que funcione de manera eficiente a bajo costo.

Introduction

Coronavirus Disease 2019 commonly known as COVID-19 is a respiratory disease mainly caused due to the SARS-COV-2 coronavirus. This virus is a kind of contagious spherical positive-sense single-stranded RNA virus, which affects the respiratory system of the body. Spikes of the protein emerging out of the surface and construct a crown-like structure can be seen with the help of

an electron microscope. Fever, cough, breathlessness, and tiredness are some of the common symptoms among the patients of COVID-19. The severity level can be range from mild to severe having chances of multi-organ failure, pneumonia, cardiovascular complications, and death in some cases. These issues arise the concern due to this pandemic due to viral infection. COVID-19 disease emerged as a major health issue this year and the number of COVID-19 affected patients are significantly increasing day by day which arises the need for countermeasures to control the spread and rapid diagnostic systems are to be developed. Thus, suitable approaches should be determined in the direction of solutions for COVID-19 related problems and extract the proper information from the large volume of data gathered associated with this.

As on 22 July, more than 14 million confirmed cases of COVID-19 cases are reported from 216 countries and other territories [1]. COVID-19 caused more 0.6 million deaths worldwide. Lots of work is going on to develop the vaccine to fight with this pandemic, but still, no specific vaccine or treatment is available. In parallel, COVID-19 cases are increasing day by day. Many attempts have been made to develop a rapid and accurate detection method to diagnose infected patients in their early stages of infection. Thus, artificial intelligence can play an important role in that direction.

In most of the clinical screening COVID-19 patients, Reverse Transcription Polymerase chain reaction (RT-PCR) is considered as a reference method and prominently used based on the analysis of respiratory samples [2]. These medical and pathological ways to detect the COVID-19 take longer time and testing facility is limited, laborious, and has high cost as well. Hence, it causes a delay in the disease prevention measures and various countries are facing difficulties with delay in test results which sometimes results in the wrong number of COVID-19 positive cases.

A lot of research is undergoing to control the spread of COVID-19 and develop screening devices. A hybrid COVID-19 detection neural network is developed where improved marine predator algorithm for segmentation of chest X-ray image [3]. Built-in smartphone sensor-based application is prepared for abnormality detection in CT scan images. An unsupervised pleural line localization and detection method from a lung ultrasound image is proposed in [4]. Support vector machines, Viterbi algorithm, and Hidden Markov model were used for evaluation of the health condition of the person.

Different machine learning methods were also analyzed for prediction of the COVID-19 affected population [5]. Exponential smoothing works best among those four machine learning models for the statistical dataset [6] of COVID-19 affected population. A contactless patient positioning system using different components such as automated positioning, calibration, and view synthesis routines is developed and robust dynamic fusion algorithm to develop 3-D model of the patient's body [7]. A secured fog-based communication architecture was designed for the timely monitoring of the patients [8]. Similarly, COVID-19 spreaders identification method is presented using the economic and socio-cultural characteristics relating to number of death and infections [9].

The main contribution of this paper is that the multisensory COVID-19 detection framework is proposed. The multi-sensor analysis consists of four steps of testing-temperature, oxygen saturation, chest X-ray, and cough sound. Different chest X-ray and cough sounds of different persons are collected having different health conditions. Two deep learning models will be trained with chest X-ray and cough sound each. The trained model after a certain number of iterations and the moment when validation loss will be lower and validation accuracy will be higher, training will be halted. The trained models will be tested on different parameters to get

the actual performance of the model. Once performance outcomes reach a satisfactory level, it will be deployed to test new images or sounds. The proposed artificial intelligence based framework can be used to separate potentially COVID-19 infected patients.

Section II will discuss the methodology of different sensors and their working to implement the COVID-19 detection framework. Section III is the conclusion and future scope is discussed.

Proposed Methodology

Relying on a single method for screening of COVID-19 cannot be the perfect solution. Some are accurate but are time-consuming at the same time also. At this time of pandemic where there is need of large-scale and fast screening, a screening framework is required which can assure the confidence in screening procedures using a multi-dimensional approach. Artificial intelligence integrated with the multiple sensors can help to deal with this problem of COVID-19 pandemic. The proposed methodology is composed of four steps, which is represented by the block diagram as shown in Fig. 1

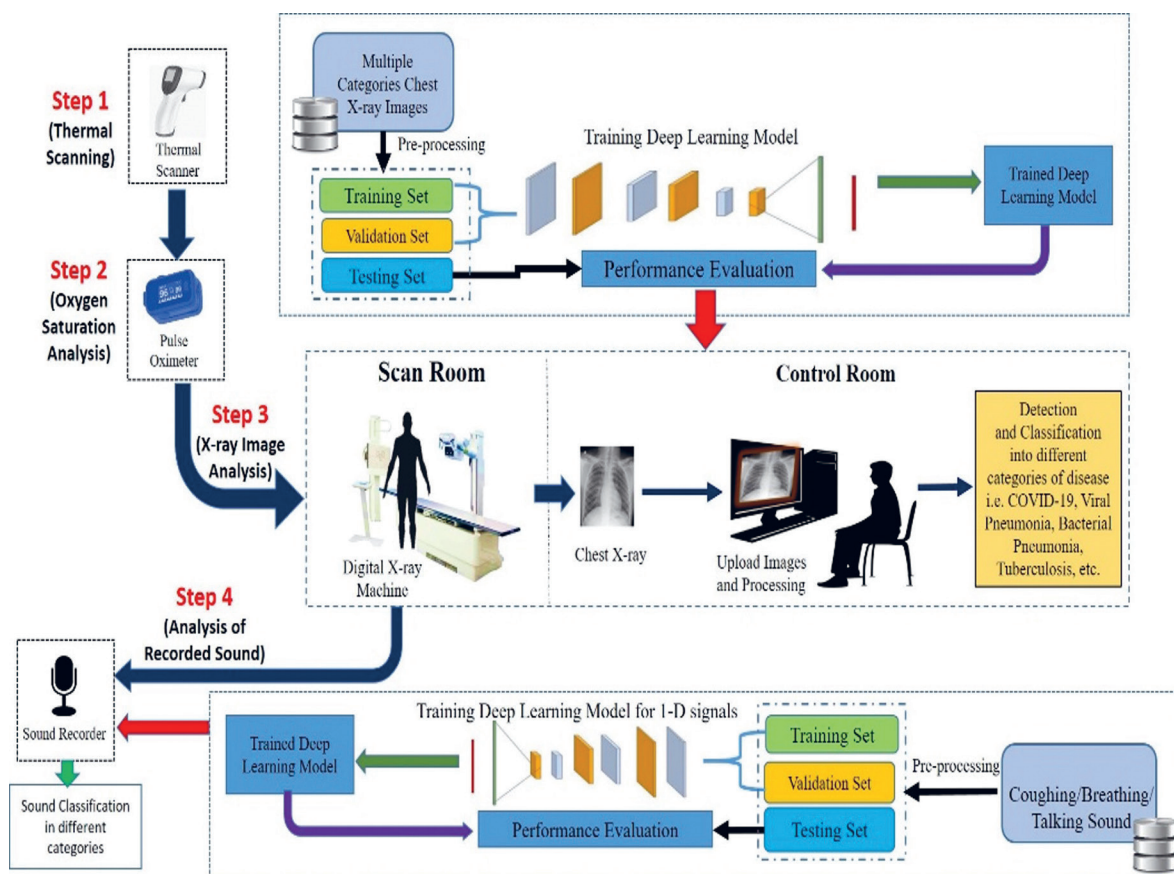


Figure 1. Proposed COVID-19 screening framework

Step-1 (Temperature Analysis)

The common symptoms of COVID-19 include the fever kind of symptoms. Therefore, to differentiate those people having abnormal temperature from normal people, a temperature scanner is to be placed at entrance of checking system. The temperature scanning can be done

manually by considering proper protective measures. The alternate way is to use the automatic temperature using the face detector based positioning of the sensor, as shown in Fig. 2. The face is searched in given input frame and once face is detected, position of servo motor attached to thermal scanner is adjusted according to the obtained coordinates of forehead.

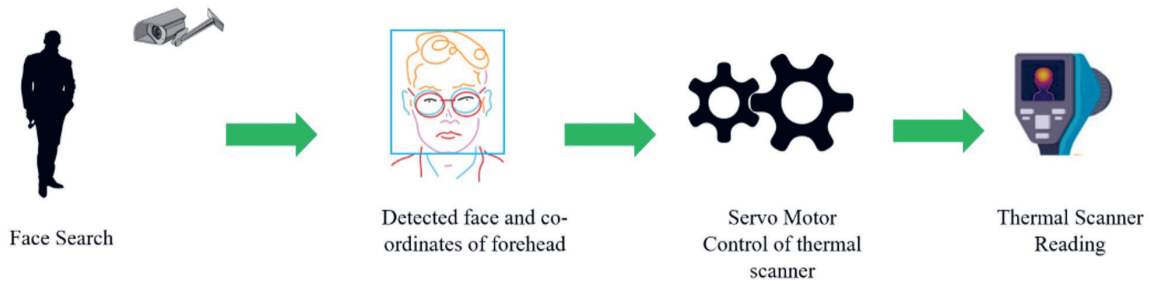


Figure 2. Face position estimation and temperature scanning.

Temperature sensing and decision equations on the basis of temperature of forehead, T_{forehead} , are given by

$$\text{Normal: } T_{\text{forehead}} \leq 99^{\circ}\text{F} \quad (1)$$

$$\text{Mild Symptom: } 99^{\circ}\text{F} < T_{\text{forehead}} \leq 100^{\circ}\text{F} \quad (2)$$

$$\text{Severe Symptom: } T_{\text{forehead}} > 100^{\circ}\text{F} \quad (3)$$

If the forehead temperature, $T_{\text{forehead}} > 99^{\circ}\text{F}$, it will warn it as mild symptom. In such cases, if the temperature goes below the 99°F or shows temperature-decaying pattern after few time or taking some rest, it will be treated as normal condition. Otherwise, it will give temperature related warning and advise for treatment and health checkup.

Step-2 (Oxygen Saturation Analysys)

An oximeter is a device, which is used to check the oxygen saturation level of the blood and can be checked while taking input from the fingertips. The pulse oximeter has seen significant advancement in the field of the clinical monitoring system. It is photometric technology-based non-invasive device that is used to measure the heart rate and blood oxygen saturation level (SpO₂). The block diagram of oximeter and its working is given in Fig. 3. The difference in the light absorbed by the tissues for two lights of different wavelengths is used to measure the oxygen saturation level of the blood. Wavelengths of light should be chosen properly such that they can give a large difference in the extinction coefficients of oxyhemoglobin and deoxyhemoglobin. Thus, red (660 nm wavelength) and near-infrared (940 nm wavelength) light are good choice for the same. Based on the study in [10], below the 92% oxygen saturation level is considered as more likely to be admitted in the hospital in an intensive care unit.

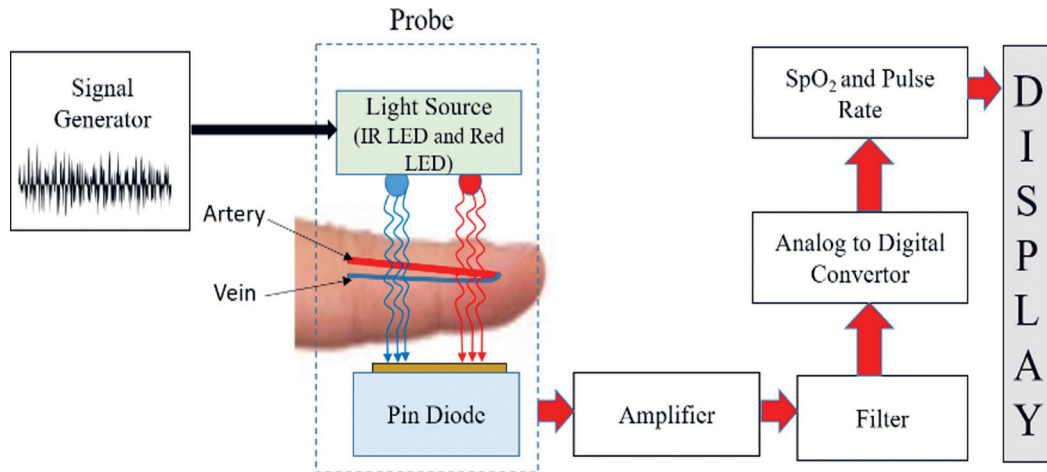


Figure 3. Pulse Oximeter components and working.

Similar to the temperature sensor, there are three condition for different oxygen saturation level, O_{sat} as given by the equations:

$$\text{Normal: } O_{sat} > 92\% \quad (4)$$

$$\text{Mild Sympton: } 90\% < O_{sat} \leq 92\% \quad (5)$$

$$\text{Critical: } O_{sat} < 90\% \quad (6)$$

Person whose oxygen saturation is below 90% need to be admitted in the hospital immediately and to be retain in proper care.

Step-3 (Chest X-ray Analysis)

As the COVID-19 is a kind of respiratory disease, many symptoms or abnormal changes can be seen in the chest X-ray of the person infected with this disease. Some asymptomatic cases where temperature and other parameters are measured as normal can be detected with the help of analysis of chest X-ray. To make such kinds of decisions and analysis expert radiologists are required. But the ratio of the expert radiologists in the larger population is very less, thus artificial intelligence-based methods made for object detection and classification such VGG-16 [11], VGG-19 [12], YOLO, Capsule Networks can be trained with the large volume of chest X-ray image data having the labels provided by the expert radiologists. The screening system can be trained with multiple classes of respiratory diseases such as viral pneumonia, bacterial pneumonia, tuberculosis and SARS for multi-classification of diseases based on chest x-ray images. In case of less availability of the labeled dataset, a chest X-ray image scan be augmented with the multiple numbers of traditional techniques such as rotation, scaling, shear, illumination variation, noise induction, etc. More accuracy can be achieved with the COVID-19 positive and negative patients, so that system can extract better features to discriminate between two classes.

The chest x-ray dataset of COVID-19 is collected from the online dataset from multiple countries [13]. 1000 number of chest X-ray images from each class of healthy and pneumonia affected people is also collected [14]. The dataset is divided in training, validation and testing set in the ration of 14:3:3 respectively, as given in Table 1.

Table 1. Dataset description of chest X-ray images.

Category	Dataset	Training Set	Validation Set	Testing Set
COVID-19	237	167	35	35
Pneumonia	1000	700	150	150
Healthy	1000	700	150	150
Total	2237	1567	335	335

The dataset is trained on two deep learning models namely, VGG-16 and VGG-19. The training parameters are as follows in Table 2.

Table 2. Training parameters for deep learning models.

Parameters	Value
Epoch	1000
Padding	'same'
Learning rate	0.001
Momentum	0.90
Batch size	32
Dropout rate	0.2-0.5

VGG-16 and VGG-19 models are trained for 1000 number of epochs and best model having lower validation loss is saved. These models are evaluated on the unseen test set of chest X-ray images. The output confusion matrix is shown in Fig. 4.

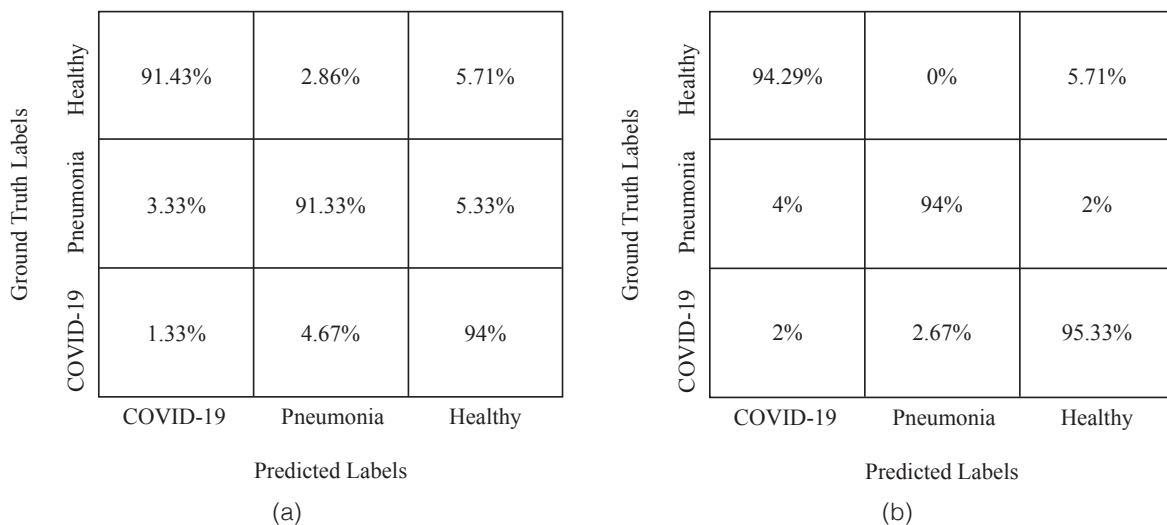


Figure 4. Confusion Matrices, (a) VGG-16, (b) VGG-19.

Confusion matrix is showing the high discrimination capability of trained model. VGG-19 model is showing good classification results for different classes. The overall classification accuracy for VGG-16 and VGG-19 is achieved as 92.53% and 96.12%, respectively.

Generative Adversarial Networks (GAN) can also be equipped to generate similar images from a limited image dataset [15]. GAN is equipped with two neural networks that compare the results with one another to generate new synthetic images that can be used to train the deep learning along with the original images, as shown in Fig. 5.

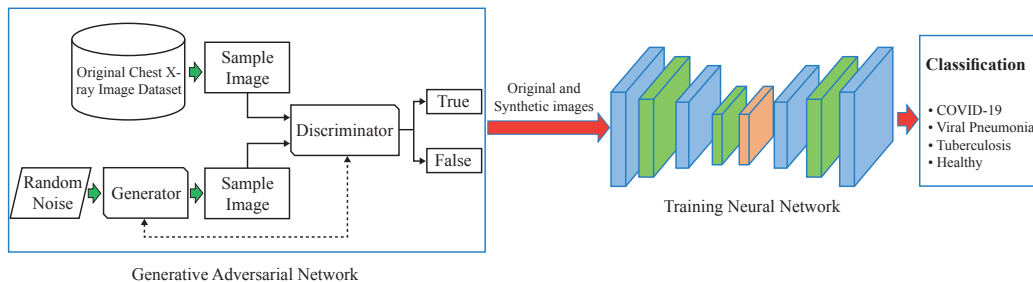


Figure 5. GAN generated trained data and training network.

Step-4 (Audio Signal Analysis)

Step 4 is an advanced step in COVID-19 screening. The main challenge is to get a proper dataset of audio samples from the people belonging to different classes i.e. healthy, tuberculosis, COVID-19, dry cough, etc. Voice or 1-D data sample will be recorded while a person is coughing, talking, breathing, etc. Anything that affects the respiratory system of the body, shows the impact in the voice of that person. These signals will be properly analyzed and the part of the signals not belonging to the required frequency range will be neglected. After having pre-processing, the signals will be used to train the convolutional neural network for multiple iterations. Training will try to extract different voice parameters and train them in different classes according to the resemblance. When the validation loss is lowered and accuracy will be intensified, training will be halted and the performance of the trained model will be analyzed based on the multiple parameters. Signal augmentation can also be applied before training of the neural network in cases where data is limited.

Conclusions

Different sensor and artificial intelligence-based framework is proposed in this paper which can be implemented for automatic and rapid screening of COVID-19. Hence, more cases can be traced easily and can be implemented anywhere at low cost with limited resources. The proposed framework test temperature, oxygen saturation level, chest X-ray, and voice samples of the person and analyzed all the samples and person having the symptoms or any abnormality can be differentiated from the healthy persons. The system is composed of four steps and all the steps analyze different parameters and all are time efficient. VGG-16 and VGG-19 models trained on chest X-ray images of different classes gives overall accuracy of 92.53% and 96.12%, respectively. The proposed framework can help to control the spread of COVID-19 by detecting the COVID-19 affected patients and verifying with multiple sensor-based methods. Portable X-ray

machine is easily available but needed a space to install, thus the installation needs to be deployed at the higher sensitive areas like airports, clinics etc. The artificially intelligent based framework can be helpful in getting faster result to diagnose potentially COVID-19 infected people.

References

- [1] World Health organization, Coronavirus disease (COVID-19) pandemic, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> (accessed on 22 July, 2020).
- [2] W. Wang, Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, et al., "Detection of SARS-CoV-2 in Different Types of Clinical Specimens," *Jama*, 2020.
- [3] M. Abdel-Basset, R. Mohamed, M. Elhoseny, R. K. Chakraborty and M. Ryan, "A Hybrid COVID-19 Detection Model Using an Improved Marine Predators Algorithm and a Ranking-Based Diversity Reduction Strategy," in *IEEE Access*, vol. 8, pp. 79521-79540, 2020. doi: 10.1109/ACCESS.2020.2990893
- [4] L. Carrer et al., "Automatic Pleural Line Extraction and COVID-19 Scoring from Lung Ultrasound Data," in *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*. doi: 10.1109/TUFFC.2020.3005512
- [5] F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in *IEEE Access*, vol. 8, pp. 101489-101499, 2020. doi: 10.1109/ACCESS.2020.2997311
- [6] Johns Hopkins University Data Repository. Cssegisanddata. Accessed: June. 27, 2020. [Online]. Available: <https://github.com/CSSEGISandData>
- [7] S. Karanam, R. Li, F. Yang, W. Hu, T. Chen and Z. Wu, "Towards Contactless Patient Positioning," in *IEEE Transactions on Medical Imaging*. doi: 10.1109/TMI.2020.2991954
- [8] C. Guo, P. Tian and K. R. Choo, "Enabling Privacy-assured Fog-based Data Aggregation in E-healthcare Systems," in *IEEE Transactions on Industrial Informatics*. doi: 10.1109/TII.2020.2995228
- [9] E. Montes-Orozco et al., "Identification of COVID-19 Spreaders Using Multiplex Networks Approach," in *IEEE Access*, vol. 8, pp. 122874-122883, 2020. doi: 10.1109/ACCESS.2020.3007726
- [10] Shah, S. et al. Novel use of home pulse oximetry monitoring in COVID-19 patients discharged from the emergency department identifies need for hospitalization. *Acad. Emerg. Med.* (2020) doi:10.1111/acem.14053.
- [11] Simonyan, K., Zisserman, A., 2014. VGG-16. *arXiv Prepr.* <https://doi.org/10.1016/j.infsof.2008.09.005>
- [12] Kim, J., Lee, J.K., Lee, K.M., 2016. Accurate image super-resolution using very deep convolutional networks, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/CVPR.2016.182>
- [13] Joseph Paul Cohen and Paul Morrison and Lan Dao COVID-19 image data collection, *arXiv*: 2003.11597, 2020 <https://github.com/ieee8023/COVID-chestxray-dataset>.
- [14] Chest X-Ray Images (Pneumonia) <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>.
- [15] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman and P. R. Pinheiro, "CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection," in *IEEE Access*, vol. 8, pp. 91916-91923, 2020, doi: 10.1109/ACCESS.2020.2994762.





A deep learning approach for epilepsy seizure detection using EEG signals

Un enfoque de aprendizaje profundo para la detección de ataques de epilepsia mediante señales de EEG

Manoj Kaushik¹, Divyanshu Singh², Malay Kishore-Dutta³, Carlos M. Travieso⁴

Kaushik, M.; Singh, D., Kishore-Dutta, M.; Travieso, C.M. A deep learning approach for epilepsy seizure detection using EEG signals. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 110-118.

 <https://doi.org/10.18845/tm.v35i8.6461>

- 1 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University. India. E-mail: manojkaushik93@gmail.com
 <https://orcid.org/0000-0002-5970-7321>
- 2 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University. India. E-mail: divyanshu0495@gmail.com
 <https://orcid.org/0000-0002-6325-5153>
- 3 Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University. India. E-mail: malaykishoredutta@gmail.com
 <https://orcid.org/0000-0003-2462-737X>
- 4 Signals and Communications Department, IDeTIC, University of Las Palmas de Gran Canaria, Las Palmas. Spain. E-mail: carlos.travieso@ulpgc.es
 <https://orcid.org/0000-0002-4621-2768>

Keywords

EEG Signal; epilepsy detection; Convolutional Neural network; pattern recognition.

Abstract

Electroencephalogram (EEG) is an effective non-invasive way to detect sudden changes in neural brain activity, which generally occurs due to excessive electric discharge in the brain cells. EEG signals could be helpful in imminent seizure prediction if the machine could detect changes in EEG patterns. In this study, we have proposed a one-dimensional Convolutional Neural network (CNN) for the automatic detection of epilepsy seizures. The automated process might be convenient in the situations where a neurologist is unavailable and also help the neurologists in proper analysis of EEG signals and case diagnosis. We have used two publicly available EEG datasets, which were collected from the two African countries, Guinea-Bissau and Nigeria. The datasets contain EEG signals of 318 subjects. We have trained and verify the performance of our model by testing it on both the datasets and obtained the highest accuracy of 82.818%.

Palabras clave

Señal EEG; detección de epilepsia; red neuronal convolucional; reconocimiento de patrones.

Resumen

El electroencefalograma (EEG) es una forma eficaz y no invasiva de detectar cambios repentinos en la actividad neuronal del cerebro, que generalmente se produce debido a una descarga eléctrica excesiva en las células cerebrales. Las señales de EEG podrían ser útiles en la predicción de convulsiones inminentes si la máquina pudiera detectar cambios en los patrones de EEG. En este estudio, hemos propuesto una red neuronal convolucional (CNN) unidimensional para la detección automática de crisis epilépticas. El proceso automático puede ser conveniente en las situaciones en las que un neurólogo no está disponible y también ayudar a los neurólogos en el análisis adecuado de las señales de EEG y el diagnóstico de casos. Hemos utilizado dos conjuntos de datos de EEG disponibles públicamente, que se recopilaron de los dos países africanos, Guinea-Bissau y Nigeria. Los conjuntos de datos contienen señales de EEG de 318 sujetos. Hemos entrenado y verificado el rendimiento de nuestro modelo probándolo en ambos conjuntos de datos y obtuvimos la precisión más alta del 82,818%.

Introduction

Epilepsy seizure is a neurological disorder. It could affect people in every age group and has been known for a very long time since 4000BC. Around 50 million people suffered from epilepsy in various parts of the world especially in economically backward countries [1]. Symptoms of epileptic seizures varies depending on the part of the brain in which electric discharge affect first. Unconsciousness, disturbance in movement, sensation and other cognitive function are some of the temporary symptoms of the disease.

Seizures with epilepsy can be categorized into two parts, behavioral and electrographic. A behavioral seizure results in physical disturbances in the body of the patient and is monitored by using the observer or video recording while the electrographic seizure is characterized by abnormal paroxysmal EEG patterns. Usually, visual analysis of EEG patterns is the method used

for epilepsy detection by the experts but this method has its own limitations including time-intensive tasks and the possibility of human error. For the above reasons, automatic detection of epilepsy has been an area of research and concern since the 1970s [2].

The automatic detection of epilepsy involves stages like recording of EEG signals from the patients, EEG signal preprocessing and analysis using different techniques such as time analysis, frequency analysis, fast Fourier transform (FFT) to extract features and finally use the extracted features to classify the epilepsy cases from the healthy ones. Various techniques have been used for EEG signals analysis such as time series analysis, frequency analysis, wavelet, fast Fourier transform, etc. For classification purposes, various machine learning and deep learning algorithms have been used like SVM, K-means, ANN, and RNN.

The main contribution of this paper is a deep learning based one dimensional convolutional neural network model which could assist a neurologist to decide whether the EEG signal is epileptic signal or control signal. It is a reliable method to identify epilepsy patterns in EEG signals and prevent errors in decision making. The proposed method is frugal, non-invasive, automatic, reliable and fast. The proposed model is trained with epileptic and control (healthy) EEG signals and able to identify epileptic signals in real time.

The rest of the paper is organized as follows: Section II discusses about the related work in the area of epilepsy detection using artificial intelligence techniques, Section III discusses about the dataset and preprocessing of raw EEG signals, Section IV contains discussion about the proposed CCN model, Section V includes discussion about the obtained results which leads to Section VI which is about conclusion and future work.

Related Work

There are various applications of EEG signals [12]. Guerra et al. [3] used wavelet transforms and neural networks to detect epilepsy seizure in EEG signals. They suggested using discrete wavelet transform (DWT) and the maximal overlap discrete wavelet transform (MODWT) to extract features from EEG signals and then use feed forward artificial neural networks (FF-ANN) for classification. They used the dataset provided by the University of Bonn consisting of five parts with each part containing 100 segments of 23.6 seconds of EEG signals. They conducted two different experiments, in the first one they selected only two sets out of five and in the second experiment all five sets were considered from the dataset. Second experiment has shown better results than first experiment [3].

Similarly, Ming yang et al. [4] extended the work further by implementing double-density discrete wavelet transform (DD-DWT) instead of traditional DWT to transform the EEG signals into sub-bands and later using genetic algorithm optimized support vector machines for classification. Input features like Hurst exponent (HE) and fuzzy entropy have been used and found to achieve even better accuracies. The dataset used by Ming yang et al. was same as that used by Guerra [3] which is openly provided by the University of Bonn. For different combinations of sets obtained results were remarkable [4].

Sathak Gupta et al. [5] have presented epilepsy seizure detection using EEG signals. The authors have proposed to use wavelet transform for coefficients extraction and from the extracted coefficients various statistical features like mean, standard deviation, root mean square, skew, kurtosis, maximum fractal length (MFL), coefficient of variation and Shannon entropy were extracted. In the next step various machine learning and deep learning algorithms were experimented on these features. The results indicate that deep learning models have better performance on pre-processed EEG signals than raw EEG signals with extreme gradient boosting [5].

Dataset and Pre-processing

The two EEG Datasets we have used and analyzed for epilepsy signal detection is openly released by Vincent van Hees and Wim Otte [6]. These datasets was created by EEG signal recordings of two African countries named as Guinea-Bissau and Nigeria. EEG signals were recorded with a low-cost, fourteen channels EEG Emotive headset. The datasets contains a total of 318 EEG signals in which 179 are of epilepsy and 139 are of control or normal signals. Country wise, dataset-A (Guinea-Bissau) contains total 97 EEG signals and dataset-B (Nigeria) contains total of 221 EEG signals. The description of datasets are shown in Table 1.

Table 1. Description of Datasets

Datasets	Seizure-Category	Number of Samples
Dataset-A (Guinea-Bissau)	Control	46
	Epilepsy	51
Dataset-B (Nigeria)	Control	93
	Epilepsy	128
Total		318

EEG signals were recorded for the three minutes of closed eyes followed two minutes of opened eyes or three minutes of opened eyes followed two minutes of closed eyes which is random among the participants. EEG signals contains 14 channels which are recorded by 14 electrodes located on the scalp as per the international 10-20 system [7]. The position of electrodes can be viewed in Fig 1. and the names of the electrodes are O1, O2, P7, P8, T7, T8, AF3, AF4, F3, F4, F7, F8, FC5, and FC6.

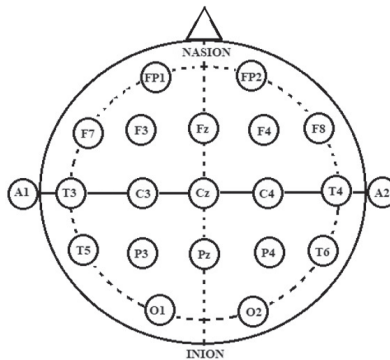


Figure 1. EEG raw data of 14 channel.

Pre-processing is done by removing the outliers from the raw EEG data. The general workflow of methodology is shown in Fig. 2. Initially all 14 channel values are arranged column wise. All the corresponding rows are deleted if any value in a row is smaller than a minimum threshold S_v or larger than a maximum threshold L_v . The values of S_v and L_v are decided based on the statistical method in which mean and standard deviation is calculated for each electrode for each row and then lower and upper bound is decided by subtracting four times the standard deviation from the pre calculated mean. Based on these bounds, we got the values S_v and L_v which is helpful in removing all of the rows for which the value is smaller or larger than these thresholds.

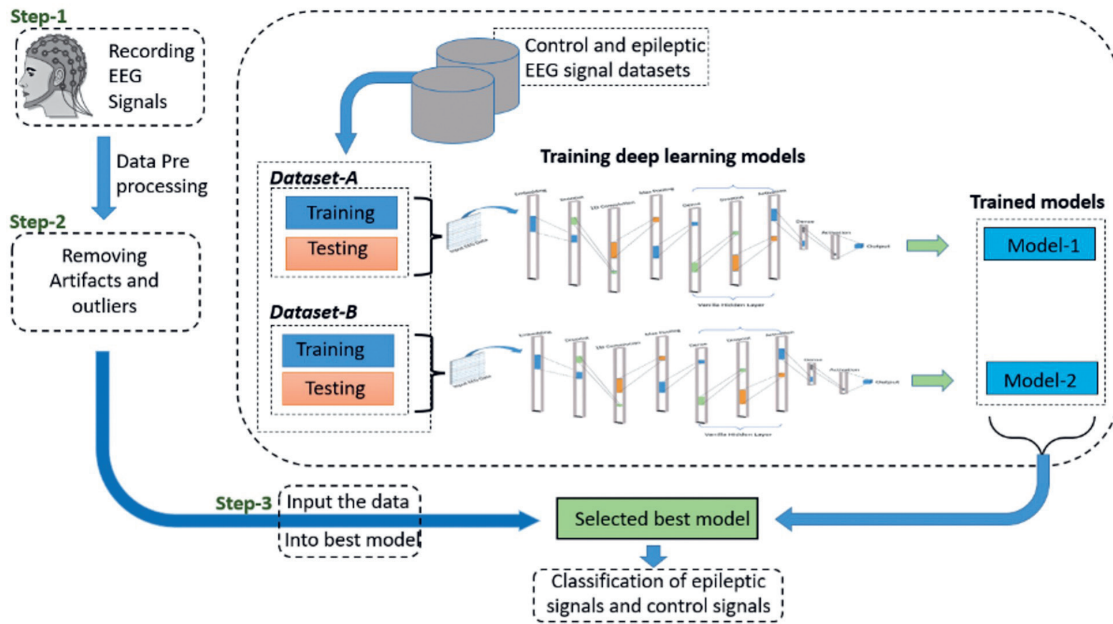


Figure 2. Workflow of the methodology.

After removal of the outliers from the data, transpose of the whole data matrix is performed. All outlier pre-processing is done on both the EEG datasets collected from Guinea-Bissau and Nigeria. Sequence padding is used before splitting the data into training and testing sets. The 14 channel EEG data after outlier removal is shown in Fig.3.

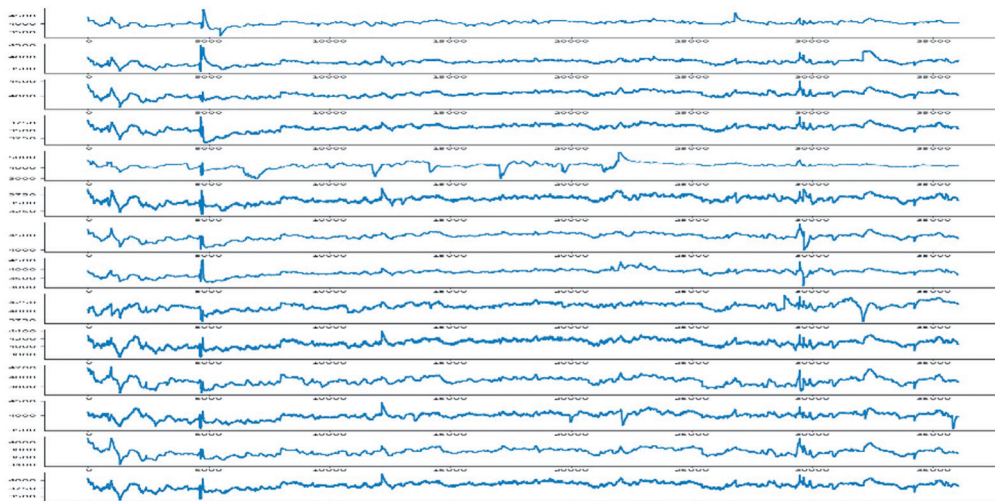


Figure 3. 14 channel EEG data after outlier removal.

Dataset-A contains a total of 97 EEG recordings while dataset-B contains a total of 221 EEG recordings. Training and testing are split by 66% and 34% respectively for both the datasets. The length of the time series EEG signals is unequal, hence equalization of all EEG signal is done by padding and truncating operations. EEG signals which are shorter than fixed length threshold L_v have padded and signals which are larger than threshold L_v have truncated.

CNN Model Architecture

We have used a 1-dimensional convolutional neural network for epilepsy and control EEG signal classification. It contains various hidden layers in the network architecture. The first layer is the embedding layer which is used to convert positive integer values into fixed-sized vectors.

After the embedding layer, the dropout layer is added to prevent the model from over fitting. The dropout layer randomly drop 20% of the neurons. Then, a 1-D convolutional layer is added in the network with 80 output filters. Output filters are used to store the extracted discriminating information by the kernel. Kernel size is set to 3, is a 1D-convolutional window which helps to extract the features.

The hyper parameter, stride value set to be 1 which specifies the step length of the convolutional window. Padding hyper parameter is valid so, no padding is used in the convolution operation. Rectified Linear Unit (ReLU) is used for neuron activation. It calculates the function $f(x) = \max(0, x)$ which provides better non linearity and convergence. After the convolutional layer, max pooling-1D layer is added which helps in reducing the number of operations for ensuing layers without losing the salient information. Then, a vanilla hidden layer is added which is a combination of dense layer, dropout layer and activation function. Dense layer preserves the information. ReLU [8] is used as an activation function. Various used hyper-parameters are shown in Table 2.

Table 2. Hyper-parameters used.

Hyper-parameter	Value
Batch Size	32
Loss	'binary cross-entropy'
Embedding Dimensions	5
Hidden Dimensions	300
Epoch	500
Filters	80
Kernel Size	3
Stride	1
Padding	'valid'
Optimization Algorithm	'Adam'

Finally, the vanilla hidden layer output is passed to a single unit output layer which is squashed with sigmoid activation [9]. Binary cross entropy used for loss. Adam optimization function used in the model architecture.

The model architecture with various parameters, layers and corresponding input sizes are shown in Fig. 4. Total number of parameters in the model architecture are 75,881. Two types of activation functions ReLU and Sigmoid have used.

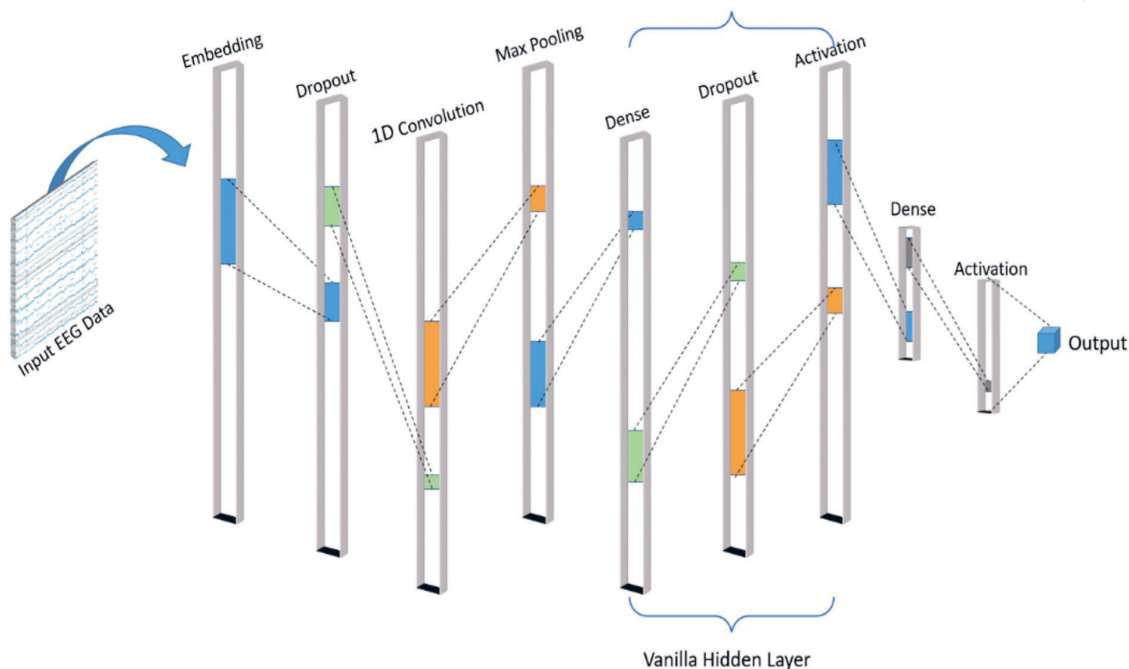


Figure 4. 14 channel EEG data after outlier removal

Results

The datasets contain EEG signals of epilepsy and control from two different countries. When we trained our model for dataset-A (Guinea-Bissau) the accuracy is 82.818% while the accuracy for dataset-B (Nigeria) is 71.493%. Dataset-A contains a total of 97 EEG signals while dataset-B contains a total of 221 EEG signals. We have also analyzed the performance of dataset-A trained model on dataset-B as a testing dataset and vice-versa.

Accuracy comparison of our proposed model and by Vincent et. el. [6] are shown in Table-3.

Table 3. Accuracy comparisons.

Trained Model on	Testing on	Accuracy (proposed model)	Accuracy By Vincent et. el
Dataset-A (Guinea-Bissau)	Test set of dataset-A	82.818%	81%
	Full dataset-B	56.023%	55%
Dataset-B (Nigeria)	Test set of dataset-B	71.493%	70%
	Full dataset-A	59.60%	60%

Accuracies for both the datasets has shown in Fig.5. Out of 33 validation sets for dataset-A 14 correctly classified as Control signals and 13 correctly classified as epilepsy while for dataset-B out of 73 validation sets 9 are correctly classified as control and 41 are correctly classified as epilepsy.

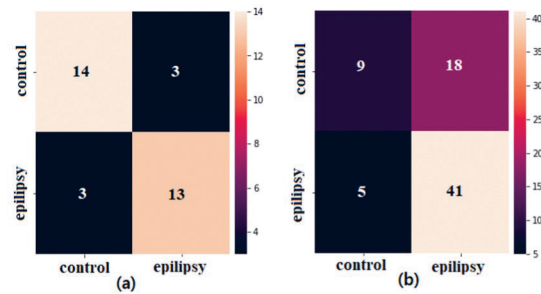


Figure 5. Confusion matrices for (a) Dataset-A (b) Dataset-B.

Conclusions

In this study, we have trained and tested a one-dimensional convolutional neural network for epilepsy and control EEG signals, collected from two African countries Guinea-Bissau and Nigeria. We have tested the trained models extensively and measured the performance on cross datasets testing. This method shows the highest accuracy performance on dataset-A (Guinea-Bissau) with 82.818%. Also features are not extracted separately instead all the features are extracted by CNN filters. Model accuracy is saturated above 500 epochs [10].

In the future different machine learning techniques with more advanced optimization algorithms can be used with separate feature extraction to get better accuracy on this dataset. One dimensional residual network (Res-CNN) can also be applied for time series data analysis [11].

References

- [1] World Health Organization Fact Sheet on Epilepsy <https://www.who.int/news-room/fact-sheets/detail/epilepsy>, latest accessed 2020/07/15.
- [2] Orosco, L., Correa, A. G., & Laciari, E. (2013). A survey of performance and techniques for automatic epilepsy detection. *Journal of Medical and Biological Engineering*, 33(6), 526-537.
- [3] Juarez-Guerra, E., Alarcon-Aquino, V., & Gomez-Gil, P. (2015). Epilepsy seizure detection in EEG signals using wavelet transforms and neural networks. In *New trends in networking, computing, E-learning, systems sciences, and engineering* (pp. 261-269). Springer, Cham.
- [4] Li, M., Chen, W., & Zhang, T. (2016). Automatic epilepsy detection using wavelet-based nonlinear analysis and optimized SVM. *Bio cybernetics and biomedical engineering*, 36(4), 708-718.
- [5] Gupta, S., Bagga, S., Maheshkar, V., & Bhatia, M. P. S. (2020, January). Detection of Epileptic Seizures using EEG Signals. In *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)* (pp. 1-5). IEEE
- [6] Vincent van Hees, & Wim Otte. (2018). EEG data collected with Emotiv device in people with epilepsy and controls in Guinea-Bissau and Nigeria (Version 1.0) [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.1252141>, latest accessed 2020/07/15
- [7] Echallier, J. F., F. Perrin, and J. Pernier. "Computer-assisted placement of electrodes on the human head." *Electroencephalography and clinical neurophysiology* 82.2 (1992): 160-163.
- [8] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," Haifa, 2010, pp. 807-814. [Online]. Available: <https://dl.acm.org/citation.cfm>, latest accessed 2020/07/15.
- [9] J. Turian, J. Bergstra, and Y. Bengio, "Quadratic features and deep architectures for chunking," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, vol. Companion Volume:, 2009, pp. 245-248. [Online]. Available: <https://dl.acm.org/citation.cfm>
- [10] Y. A. LeCun, L. Bottou, G. B. Orr, K.-R. Muller, Efficient backprop, "in: *Neural networks: Tricks of the trade*, Springer, 2012, pp. 9-48.

- [11] D. Lu and J. Triesch, "Residual deep convolutional neural network for eeg signal classification in epilepsy," arXiv preprint arXiv:1903.08100, 2019.
- [12] Baghel, N., Singh, D., Dutta, M. K., Burget, R., & Myska, V. (2020, July). Truth Identification from EEG Signal by using Convolution neural network: Lie Detection. In 2020 43rd International Conference on Telecommunications and Signal Processing (TSP) (pp. 550-553). IEEE.

Comparison of four classifiers for speech-music discrimination: a first case study for costa rican radio broadcasting


Comparación de cuatro clasificadores para la discriminación de voz y música: un primer estudio de caso para la radiodifusión costarricense

Joseline Sánchez-Solís¹, Marvin Coto-Jiménez²

Sánchez-Solís, J.; Coto-Jiménez, M. Comparison of four classifiers for speech-music discrimination: a first case study for costa rican radio broadcasting. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 119-127.

 <https://doi.org/10.18845/tm.v35i8.6463>

1 University of Costa Rica. Costa Rica. Costa Rica. E-mail: joseline.sanchez@ucr.ac.cr

2 University of Costa Rica. Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr
 <https://orcid.org/0000-0002-6833-9938>



Keywords

Classification; music; radio broadcasting; speech.

Abstract

During the past decades, a vast amount of audio data has become available in most languages and regions of the world. The efficient organization and manipulation of this data are important for tasks such as data classification, searching for information, diarization among many others, but also can be relevant for building corpora for training models for automatic speech recognition or building speech synthesis systems. Several of those tasks require extensive testing and data for specific languages and accents, especially when the development of communication systems with machines is a goal. In this work, we explore the application of several classifiers for the task of discriminating speech and music in Costa Rican radio broadcast. This discrimination is a first task in the exploration of a large corpus, to determine whether or not the available information is useful for particular research areas. The main contribution of this exploratory work is the general procedure and selection of algorithms for the Costa Rican radio corpus, which can lead to the extensive use of this source of data in many own applications and systems.

Palabras clave

Clasificación; música; radiodifusión; habla.

Resumen

Durante las últimas décadas, una gran cantidad de datos de audio ha estado disponible en la mayoría de los idiomas y regiones del mundo. La organización y manipulación eficiente de estos datos son importantes para tareas como clasificación de datos, búsqueda de información, diarización entre muchas otras, pero también pueden ser relevantes para construir corpus para modelos de entrenamiento para reconocimiento automático de voz o construir sistemas de síntesis de voz. Varias de esas tareas requieren pruebas y datos exhaustivos para idiomas y acentos específicos, especialmente cuando el objetivo es el desarrollo de sistemas de comunicación con máquinas. En este trabajo, exploramos la aplicación de varios clasificadores para la tarea de discriminar el habla y la música en la radiodifusión costarricense. Esta discriminación es una primera tarea en la exploración de un gran corpus, para determinar si la información disponible es útil o no para áreas de investigación particulares. El principal aporte de este trabajo exploratorio es el procedimiento general y la selección de algoritmos para el corpus de radio costarricense, lo que puede llevar al uso extensivo de esta fuente de datos en muchas aplicaciones y sistemas propios.

Introduction

In our days, there is a vast amount of multimedia data, such as images, audio, and video, which are available on the Internet, radio and television broadcasts. The amount of information of this kind during the last years has seen an exponential growth [1]. The manipulation and organization of this data are required in many tasks, for example, in classification for storage, summarizing and describing the content. A large portion of the data is audio, from resources such as broadcasting radio, audio-books, internet streams, and commercial music recordings.

Due to massive amounts of this data it is impossible to generate classes, labels, descriptions, transcriptions manually, or many of the main tasks that are required to take advantage of the information [2]. To answer the demands for handling the data, a field of research, known as audio

content analysis (ACA), or machine listening, has recently emerged [1]. The purpose of ACA can be established as extracting information directly from the acoustic signal and automatically create descriptions, detect types of content, semantic annotation, speaker diarization or other tasks according to specific requirements.

One of the first tasks that need to be solved in ACA applications is the automatic distinction between music and speech. This is, due to the very different content and applications of both classes, such as genre classification (in music) or automatic speech recognition. A broadcast audio content can be annotated as music or speech once input audio is classified as speech/music segment [3]. This problem, with only two classes has been of interest in academia, and also has industrial applications, for example general purpose audio codecs [4].

Among the main challenges in performing speech/music classification is to obtain high accuracy with characteristics of short-time delay and low complexity [5], due to the vast amount of information that need to be processed and its efficient utilization in the applications. For discriminating the audio content in categories such as music and speech, data two successive stages have to be performed: i) the extraction of features from the input audio data, and ii) the classification of the data into established categories [2].

And beyond classification and semantic annotation of information, speech/music discrimination is also an essential part of speech coding, to efficiently utilize the bandwidth resources. For example, different bit rate allocations for different input formats can be applied according to the content [6]. If an automatic speech recognition system is applied to broadcast news, it is also desirable to consider disabling the input to the speech recognizer during the non-speech portion of the audio stream [7].

In this paper, we compare the performance of several algorithms for speech/music classification for a Costa Rican radio broadcasting, using feature vectors extracted from the audio [8]. To our knowledge, this is the first report for this task with Costa Rican data, which is relevant due to the dependency on many kinds

of music and speech-related tasks in the language and particular accents.

The rest of this paper is organized as follows: Section 2 presents a review of the literature related to music and speech classification. Section 3 presents the Experimental Setup used in the research. The results and discussion are presented in Section 4. Finally, the conclusions are found in Section 5.

Related Work

For discrimination of speech and music in audio signals, numerous techniques exploring several features and classifiers have been proposed. It is also important to consider that there could be a dependency on the type of recordings (radio or television broadcasts, audio quality, audio books, among others), as well as the type of music and language.

Some of the most common features include speech-specific parameters, such as zero-crossing rate (ZCR), spectral centroid, spectral roll-off, and Mel frequency cepstral coefficients (MFCC) [3]. One of the first works, presented in [9] applied a calculation of ZCR and its statistics in short segments of FM audio broadcasting, obtaining error rates below 10%.

The fixed-size set of features for audio segments is also known as i-vector, as reported in [10]. There, an analysis of algorithms such as cosine distance score (CDS), support vector machine (SVM), and linear discriminant analysis (LDA) is performed for speech and music classification in the English language, with a database derived from TIMIT.

Among the most recent features applied to this tasks, [11] includes some more complex features derived from speech analysis, such as the normalized auto-correlation peak strength (NAPS), the zero frequency filtered signal (ZFFS), the peak-to-sidelobe ratio (PSR), and Hilbert envelope (HE) of linear prediction(LP) coefficients. It is important to remark that these more complex features are applied for languages and databases where previous experimentation has been done.

A measure based on energy, called Minimum Energy Density (MED), was applied to the binary speech/music classification in the radio broadcasts in [12], for the Polish language. The primary motivation of the study was to provide information about the contents of the radio stations.

The possibility of discriminate segments of speech and music can also be approached from short segments. For example, in [4], segments in the order of tens of milliseconds are explored. The authors adopted a statistical approach, and the features were found using an unsupervised procedure. Also, in [13], a two- step segmentation approach is employed to identify transition points between homogeneous regions, and then successfully apply the algorithms for classifying the segments.

Some of the best results in the task of binary speech/music classification have been obtained using machine learning techniques, such as the support vector machine (SVM), Gaussian mixture model (GMM), and deep belief networks (DBN) [6]. In most of the references, a comparative study is performed using several features or algorithms for classification, in particular when a new set of features is proposed. Also, it is likely to experiment on databases made from clean speech and pure music, leaving aside some interesting cases, such as music plus background music.

More recently [3], experimented with chromagram-based music-specific features, claiming that such features outperform other existing approaches in terms of classification accuracy. In [14] some new algorithms, based on deep learning, also increase the accuracy of the classification.

Experimental Setup

Database used

The tests were made using a music/speech database created for this work, using audio recordings from "Comunidad 870" podcasts, a program that belongs to "Radio 870" of the University of Costa Rica. The podcasts are publicly available at <https://radios.ucr.ac.cr/>.

From 10 hours of audio recordings, an automatic segmentation based on detection of silence, with the SoX program was performed to produce 178 music files and 773 clean speech files. The resulting files do not have a standard duration. All files have different duration, with some lasting several minutes while others only seconds. 10% of the total music or speech files were used for the test files. All the files were manually classified as clean speech or music to establish the ground-truth for the classifiers.

Due that the main purpose of this exploration is the assessment of the audio resources of this podcast for further research in speech technologies, the speech files with background music were considered as music segments.

Feature extraction

The sound features used to train the classifiers were selected upon bibliographic criteria. Each set of features was tested alone and in combination with the others in every classifier, with the aim of finding the best set of features and in case of similitude, those with the simplest or least amount of features. The features were extracted using pyAudioAnalysis [15], and are described within each of the following groups:

1. **Spectral and Energy Features.** The mean value of the following features was obtained for each audio file containing music of speech:
 - Zero Crossing Rate: The rate of sign-changes of the signal during the duration of a particular frame.
 - Energy: The sum of squares of the signal values.
 - Entropy of Energy: The entropy of the sub-frames, which can be interpreted as a measure of abrupt changes.
 - Spectral Centroid: The center of gravity of the spectrum.
 - Spectral Spread: The second central moment of the spectrum.
 - Spectral Entropy: Entropy of the normalized spectral energies.
 - Spectral Flux: The squared difference between the normalized magnitudes of the spectra of the two successive frames.
 - Spectral Rolloff: The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
2. **MFCC.** The Mel Frequency Cepstrum Coefficients (MFCC) are a set of the Discrete Cosine Transform parameters, computed using a perceptually spaced triangular filter bank that processes the Discrete Fourier Transform of the speech signal [16]. Every frame of speech of music signal can be expressed as a set of MFCC coefficients. In our work, we used the mean value of each of 13 coefficients in each audio file.
3. **Chroma Vectors.** Chroma features are twelve-element vectors, where each dimension represents the intensity associated with a particular frequency of tempered musical scale; thus, each element can be associated with a semitone of the musical scale, regardless of octave [17].

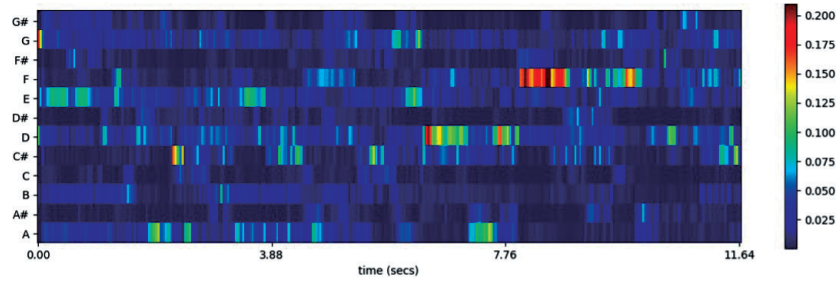
Chroma vectors have been applied successfully in music gender classification and several tasks related to music identification. In our work, we used the mean values of the twelve chroma dimensions for each audio file.

Classifiers

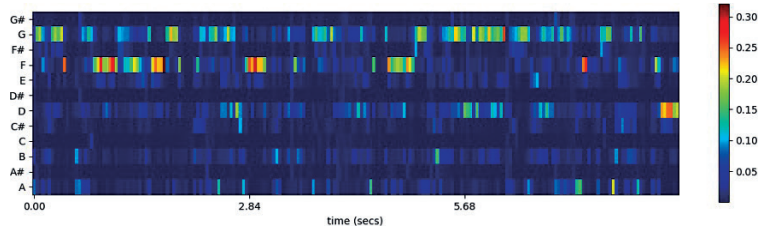
For this case study, we chose four classifiers commonly applied in exploratory studies: 1) Support Vector Machines (SVM) with a polikernel implementation. 2) K-nearest neighbors (KNN). 3) Random Forest with a number of trees of 100 and seed = 1. 4) Naive Bayes.

Results and Discussion

In this section, we present the results of the four classifiers and the combinations of features. Table 1 shows the results of the test music/speech file classification, for each of the four algorithms applied and all the subsets of features available.



(a) Music file



(b) Speech file

Figure 1. Chroma vector for two files of the database.

The purpose of this comparison is to find the best algorithm and ideally the simpler set of features that perform properly for this problem.

It can be seen that the best results are obtained with KNN, since it presents error rates as low as 1.9%, with the combinations of features Energy + MFCC, MFCC + Chroma and Energy + MFCC + Chroma. The Random Forest classifier was the classifier with the second-best results, with the Energy + MFCC + Chroma features.

Table 1. Evaluation of results for each classification algorithm and set of features.* Is the best result.

KNN				
Features	Error rate	Precision	Recall	F1 score
Energy	11.8812	0.876	0.881	0.874
MFCC	2.9701	0.927	0.981*	0.969
Chroma	8.9109	0.909	0.911	0.907
Energy+MFCC	1.9802*	0.981*	0.980	0.980*
Energy+Chroma	4.9505	0.950	0.950	0.950
MFCC+Chroma	1.9802*	0.981*	0.980	0.980*
Energy+MFCC+Chroma	1.9802*	0.981*	0.980	0.980*
SVM				
Features	Error rate	Precision	Recall	F1 score
Energy	11.8812	0.897	0.881	0.863
MFCC	14.8515	0.857	0.851	0.826
Chroma	21.7822	0.612	0.782	0.687
Energy+MFCC	10.8911*	0.904*	0.891*	0.876*
Energy+Chroma	10.8911*	0.904*	0.891*	0.876*
MFCC+Chroma	13.8614	0.882	0.861	0.831
Energy+MFCC+Chroma	10.8911*	0.904*	0.891*	0.876*
Random Forest				
Features	Error rate	Precision	Recall	F1 score
Energy	9.901	0.899	0.901	0.895
MFCC	4.9505	0.950	0.950	0.949
Chroma	8.9109	0.920	0.911	0.902
Energy+MFCC	3.9604	0.962	0.960	0.956
Energy+Chroma	4.9505	0.953	0.950	0.948
MFCC+Chroma	3.9604	0.962	0.960	0.959
Energy+MFCC+Chroma	2.9703*	0.971*	0.970*	0.969*
Naive Bayes				
Features	Error rate	Preision	Recall	F1 score
Energy	18.8119	0.792	0.812	0.793
MFCC	20.7921	0.803	0.792	0.797
Chroma	22.7723	0.726	0.772	0.733
Energy+MFCC	16.8317*	0.829*	0.832*	0.830*
Energy+Chroma	21.7822	0.769	0.782	0.774

In Figure 1, the speech chromogram shows a region where the highest intensity frequencies are more concentrated, unlike music, where a wider diversity of frequencies are highlighted. Even though this visually evident differentiation, the Chroma-vector isolated parameters are not performing properly, compared to the combination of features.



According to the results, the best combination of features for this problem in the dataset developed is Energy + MFCC + Chroma, which presents a lower error rate in three of the four classifiers used. No specific combination of features presents the lower error rates for all classifiers. But on the other hand, it is clear from the results that the Energy and Chroma features are not efficient for this problem when used individually.

Other features such as individual MFCC presents error percentages of less than 10% for the KNN classifier, but for the rest classifiers, the error rates are greater than 10%.

In the case of the MFCC + Chroma features, the results are the best for the KNN classifier. However, for the Naive Bayes classifier, the percentage of error and other indexes are the most inefficient. SVM present error greater than 10%, but this is not the worst result for this classifier. For Random Forest the error percentage is 3.9604%, but it is not the most functional feature for this classifier.

Conclusions

In this work, we developed a dataset and performed the first case study of Speech/Music classification in a Costa Rican radio broadcasting. Our purpose is to evaluate the applicability of such source of audio information for research in speech technologies development, where a source of clean speech in particular languages and accent is mandatory.

From the results of four classifiers, the best results were obtained with the simpler one: KNN. Additionally, in regards to the comparison of classifiers, we performed a comparison of features, and the simplest combination with the best results corresponded to Energy+MFCC and MFCC+Chroma vector features.

When implementing an individual set of features (such as only MFCC or only Chroma-vector features), the classifiers dropped its capacity to discriminate clean speech from music or speech with background music.

For future work, we intend to implement a live broadcasting discriminator for clean speech, and build databases of Costa Rican Spanish that help the development of speech technologies for this particular Spanish accent.

References

- [1] Lavner, Yizhar, and Dima Ruinskiy. "A decision-tree-based algorithm for speech/music classification and segmentation." *EURASIP Journal on Audio, Speech, and Music Processing* 2009 (2009): 1-14.
- [2] Ghosal, Arijit, and Suchibrota Dutta. "Speech/music discrimination using perceptual feature." *Computational Science and Engineering: Proceedings of the International Conference on Computational Science and Engineering (Beliaghata, Kolkata, India, 4-6 October 2016)*. CRC Press, 2016.
- [3] Birajdar, Gajanan K., and Mukesh D. Patil. "Speech/music classification using visual and spectral chromagram features." *Journal of Ambient Intelligence and Humanized Computing* 11.1 (2020): 329-347.
- [4] Hirvonen, Toni. "Speech/music classification of short audio segments." *2014 IEEE International Symposium on Multimedia*. IEEE, 2014.
- [5] Wu, Qiong, et al. "A combination of data mining method with decision trees building for Speech/Music discrimination." *Computer Speech & Language* 24.2 (2010): 257-272.
- [6] Kang, Sang-Ick, and Sangmin Lee. "Improvement of Speech/Music Classification for 3GPP EVS Based on LSTM." *Symmetry* 10.11 (2018): 605.
- [7] Ruiz-Reyes, Nicolas, et al. "New speech/music discrimination approach based on fundamental frequency estimation." *Multimedia Tools and Applications* 41.2 (2009): 253-286.

- [8] Kim, S. B., and S. M. Lee. "A Comparative Evaluation of Speech-Music Classification Algorithms in the Noise Environment." *International Journal of Design, Analysis and Tools for Integrated Circuits and Systems* 8.1 (2019): 36-37.
- [9] Saunders, John. "Real-time discrimination of broadcast speech/music." *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 2. IEEE, 1996.
- [10] Zhang, Hao, et al. "Application of i-vector in speech and music classification." *2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2016.
- [11] Khonglah, Barriskhem K., and SR Mahadeva Prasanna. "Speech/music classification using speech-specific features." *Digital Signal Processing* 48 (2016): 71-83.
- [12] Kacprzak, Stanislaw, B-laz'ej Chwie'cko, and Bartosz Zi'o-lko. "Speech/music discrimination for analysis of radio stations." *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2017.
- [13] Tsipas, Nikolaos, et al. "Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination." *Multimedia Tools and Applications* 76.24 (2017): 25603-25621.
- [14] Li, Zhitong, et al. "Optimization of EVS speech/music classifier based on deep learning." *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE, 2018.
- [15] Giannakopoulos, Theodoros. "pyaudioanalysis: An open-source python library for audio signal analysis." *PloS one* 10.12 (2015).
- [16] Hossan, Md Afzal, Sheeraz Memon, and Mark A. Gregory. "A novel approach for MFCC feature extraction." *2010 4th International Conference on Signal Processing and Communication Systems*. IEEE, 2010.
- [17] Ellis, Daniel PW. "Classifying music audio with timbral and chroma features." (2007): 339-340.

Application of Fischer semi discriminant analysis for speaker diarization in costarican radio broadcasts

Aplicación del análisis semi discriminante de Fischer para la diarización de locutores en transmisiones de radio costarricenses


Roberto Sánchez-Cárdenas¹, Marvin Coto-Jiménez²

Sánchez-Cárdenas, R.; Coto-Jiménez, M. Application of Fischer semi discriminant analysis for speaker diarization in costarican radio broadcasts. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 128-136.

 <https://doi.org/10.18845/tm.v35i8.6464>

¹ University of Costa Rica. Costa Rica. E-mail: roberto.sanchezcardenas@ucr.ac.cr

² University of Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr

 <https://orcid.org/0000-0002-6833-9938>

Keywords

Broadcasting; clustering; speaker diarization; speech technologies.

Abstract

Automatic segmentation and classification of audio streams is a challenging problem, with many applications, such as indexing multimedia digital libraries, information retrieving, and the building of speech corpus (or spoken corpus) for particular languages and accents. Those corpus is a database of speech audio files and the corresponding text transcriptions. Among the several steps and tasks required for any of those applications, the speaker diarization is one of the most relevant, because it pretends to find boundaries in the audio recordings according to who speaks in each fragment. Speaker diarization can be performed in a supervised or unsupervised way and is commonly applied in audios consisting of pure speech. In this work, a first annotated dataset and analysis of speaker diarization for Costa Rican radio broadcasting is performed, using two approaches: a classic one based on k-means clustering, and the more recent Fischer Semi Discriminant. We chose publicly available radio broadcast and decided to compare those systems' applicability in the complete audio files, which also contains some segments of music and challenging acoustic conditions. Results show a dependency on the results according to the number of speakers in each broadcast, especially in the average cluster purity. The results also show the necessity of further exploration and combining with other classification and segmentation algorithms to better extract useful information from the dataset and allow further development of speech corpus.

Palabras clave

Radiodifusión; agrupación; registro de locutores; tecnologías del habla.

Resumen

La segmentación y clasificación automática de transmisiones de audio es un problema desafiante, con muchas aplicaciones, como la indexación de bibliotecas digitales multimedia, la recuperación de información y la construcción de corpus de voz (o corpus hablado) para idiomas y acentos particulares. Ese corpus es una base de datos de archivos de audio de voz y las transcripciones de texto correspondientes. Entre los varios pasos y tareas requeridos para cualquiera de esas aplicaciones, la diarización del hablante es una de las más relevantes, porque pretende encontrar límites en las grabaciones de audio según quién habla en cada fragmento. La diarización del hablante se puede realizar de forma supervisada o no supervisada y se aplica comúnmente en audios que consisten en habla pura. En este trabajo, se realiza un primer conjunto de datos anotados y análisis de la diarización de locutores para la radiodifusión de Costa Rica, utilizando dos enfoques: uno clásico basado en la agrupación de k-medias y el más reciente Fischer Semi Discriminant. Elegimos la transmisión de radio disponible públicamente y decidimos comparar la aplicabilidad de esos sistemas en los archivos de audio completos, que también contienen algunos segmentos de música y condiciones acústicas desafiantes. Los resultados muestran una dependencia de los resultados de acuerdo con el número de hablantes en cada transmisión, especialmente en la pureza promedio del clúster. Los resultados también muestran la necesidad de una mayor exploración y combinación con otros algoritmos de clasificación y segmentación para extraer mejor información útil del conjunto de datos y permitir un mayor desarrollo del corpus del habla.

Introduction

Speaker diarization is a process that involves the segmentation and clustering of audio recordings, to determine when each of the participants in the recording speaks. It can also be described as the process of partitioning an input audio stream into homogeneous segments according to speaker identity [1].

Due to the development and availability of massive audio-visual data during the past few years, the efficient management of audio content is becoming inevitable [2]. There is a need to implement techniques to process the video and audio data automatically, and speaker diarization is one of such leading techniques that can allow the clustering and characterization of audio information in terms of “who speak when.”

The process is also useful for many speech processing technologies which assume the presence of only one speaker, such as automatic speaker identification. For these processes, speaker diarization can be an essential front end where the single-speaker assumption cannot be considered [3].

The process involved in performing speaker diarization can be described as the clustering of segments of the same acoustic nature. This has been implemented to detect segments of the same speaker, but theoretically it can be also applied to other kind of sounds. For specific cases and conditions, it is expected that the different algorithms available perform distinctly.

In radio broadcast signals, diarization can be part of a system that detects audio parts that contain speech, music, silence and other types of sounds. Then, each part detected can be processed by speech recognizers, language recognizers, singer recognizers, song recognizers, etc. [4].

The speech corpus developed for this paper consists of the annotation of the periods of time where a labeled speaker participated in the audio stream. For the building of speech corpus that can be part of speech recognition systems or speech synthesis systems, the automatic characterization of broadcasts, in terms of longitude of speech (or music) segments, can be of great interest. This is due to the difficulties and long processes it takes to build a corpus using studio recordings.

This is the case of the present paper, where the development of speech technologies suitable for Costa Rican Spanish requires the building of large corpus of speech recordings and its corresponding transcription and labeling. The use of radio broadcasts for such research has been widely explored for several languages [5–8].

The rest of this paper is organized as follows: Section 2 provides a brief overview about Related work and the Fisher Semi-Discriminant Analysis. Section 3 describes the Experimental Setup. Section 4 presents the Results and discussion, and finally, Section 5 presents the conclusions.

Background

Related work

The building of corpus for research in speech technologies using radio broadcast have been presented in [9], for an Italian Broadcast News Corpus. There are many tasks required to segment, annotate and cluster speech from a such source of audio data. Speaker diarization is one of the first and most important of those tasks.

In [1] the diarization of broadcast recordings is based on an audio partitioner, which provides high cluster purity (99% for the best combination of features). The experiments were conducted on US English and French data from broadcast news.

More recently, the Fisher Semi-Discriminant Analysis (FLSD), developed and presented in [10] have been applied successfully for this task. In its first report, using a large corpus of 70 debate with 190 participants, using the Canal 9 Database. A refinement of the proposal was presented in [11], using different longitude of the segments in the process of clustering, providing relative improvement than the previous one.

FLSD has also been combined with visual features (such as face expression) to improve results [12]. In our work, we build an initial database of Costa Rican radio broadcast for the experimentation with the FLSD. In our case, only the audio recordings are available, so the combination with visual features is not possible. We chose to perform this first experimentation with the raw podcast as the source recordings for diarization. This also means short segments of music, advertisements and other factors are present, which represents new challenges for the algorithm.

Fisher Semi-Discriminant Analysis

FLSD is an extension of the Fisher linear discriminant analysis (FLD) method, which is used in classification in a similar way to other dimensionality reduction techniques (e.g. Principal Component Analysis), but with the clustering guided by the information from an existing mapping between linear combination of features, x with a corresponding set of classes c_k [12].

For this purpose, a set of scatter matrices are calculated from between class and within class vectors. The purpose of FLD is to perform a reduction where the class-means are well separated, measured relative to the data assigned to a particular class [13].

The advantages of FLD relies on the employ of information from the classes, but this advantage cannot be considered in unsupervised cases, such as speaker diarization. For this reason, FLSD, presented in [10] only requires a set of features vectors that belong to the same class, given any feature vector.

In temporal data such as an audio recording, it is expected that given any sample of the audio, the neighbor samples most likely belong to the same speaker or music segment. With this information it is possible to estimate the scatter matrices required in FLSD closer to those of the case where the original classes are known. Further details of this technique can be found in [12, 10]. In our work, we use the implementation presented in [14].

Experimental setup

Database

In order to evaluate the performance of the Fisher Semi-Discriminant Analysis in a Spanish speaking database, a new corpus was developed. This corpus is based on the Costa Rican radio program obtained from Radio Universidad, a broadcast from the University of Costa Rica. The program used for this corpus is named *Desayunos* (Breakfasts), it has different hosts, guests and ads every day, which makes it an ideal program for the evaluation of the algorithm.

A total of ten complete programs were used and divided in half to obtain more accurate results. Every program was annotated in a tab-separated value format.

The test runs where made on raw audio, which contained speakers, silence, music and other sounds from the ads. Most programs were recorded using an internet communication stream, which caused quality losses in the audio in small periods.

Evaluation

As stated in [10], a set of two common metrics is provided by FLSD method: Average Speaker Purity (ASP) and Average Cluster Purity (ACP). Both measurements require a ground truth of annotated speaker turns in the audio files.

The cluster purity measures the frequency of the most common speaker into each cluster. The higher the ACP means higher information of a unique speaker within each cluster. Speaker purity is based on the frequency of the most common detected speaker within each speaker class. The higher the ASP means a better coincidence of the detected speaker turns.

Results and Discussion

Two main tests were made to the data; with a known and an unknown number of speakers in a specific program. This was made in order to prove the capabilities of the software to be used in supervised and unsupervised diarization. Due to the annotations a percentage of cluster and speaker purity is obtained. Every program was tested on the basic speaker diarization algorithm and FLSD.

Table 1 presents the result from test runs where the number of speakers wasn't specified beforehand. The results of the base system (based on K-means) and FLSD are shown side by side in order to compare them. The ACP variate in a range from 37.6% to 98% using the base system, while FLSD has a range from 49.7% to 100%. FLSD improved the diarization up to 16.7% and had no significant worsening. Overall, FLSD improved cluster purity by 3.2%, where there is an unknown number of speakers. The standard deviation was reduced by 14.34 using FLSD, which means it associates speakers more accurately.

As for speaker purity, the base system had a range from 26.9% to 98.5%, while FLSD varies from 54.6% to 99.9%. Overall, FLSD improved speaker purity by 8,4% and had a standard deviation of 13.02 compared to 19.26 using the base system.

It can be seen in Figure 1, that with an unknown number of speakers, the speaker purity is almost constant. Cluster purity tends to decrease significantly as there are more speakers and it is noticeable that both FLSD and simple method perform similarly.

Table 1. Cluster and speaker purity. Unknown number of speakers.

Program	Base system		FLSD	
	ACP (%)	ASP (%)	ACP (%)	ASP (%)
1	37.6	72.7	54.3	
2	68.1	73.2	79.4	98.0
3	86.5	97.6	87.1	97.7
4	86.0	97.9	86.1	98.0
5	48.0	88.1	50.1	82.5
6	43.0	84.6	49.7	93.1
7	65.0	99.6	64.9	99.6
8	62.1	99.0	62.0	99.2
9	76.2	82.7	78.6	99.9
10	82.0	62.0	85.2	99.9
11	75.2	92.7	77.7	95.4
12	74.9	97.7	74.8	97.2
13	66.4	79.5	66.8	97.7
14	59.6	94.4	70.3	97.3
15	65.4	97.6	65.4	97.7
16	74.1	98.5	74.3	98.9
17	81.8	95.5	82.9	95.5
18	74.2	87.5	79.6	88.0
19	98.0	26.9	99.2	57.8
20	97.5	48.4	100	54.6
MEAN & SD	71.08 (16.14)	83.80 (19.26)	74.28 (14.34)	92.20 (13.02)

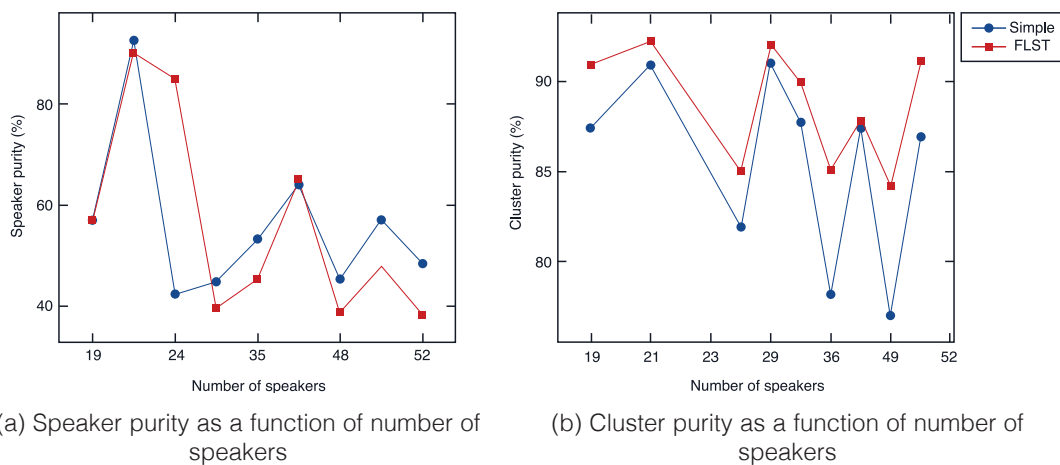


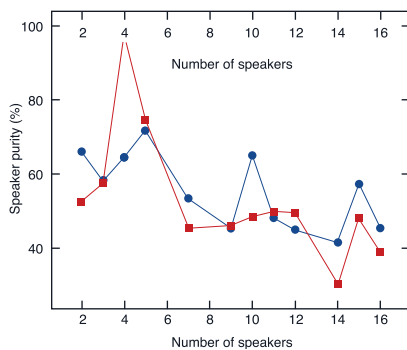
Figure 1. Unknown number of speakers ACP and ASP results.

as there are more speakers and it is noticeable that both FLSD and simple method perform similarly.

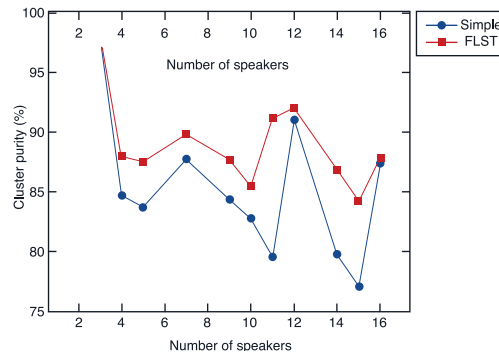
Supervised diarization results are presented in Table 2. With this method cluster purity varies in a range from 70% to 98% in the base system, while FLSD varies from 82.6% to 99.2%. On average the results were improved by 5,06%. ACP using FLSD had an average of 90.17% and a standard deviation of 4.63. There's an overall improvement using the FLSD method compared with the base method.

Table 2. Cluster and speaker purity. Fixed number of speakers.

Program	Base system		FLSD	
	ACP (%)	ASP (%)	ACP (%)	ASP (%)
1	87.4	45.3	87.8	38.7
2	91.0	44.8	92.0	39.5
3	86.9	48.4	91.1	38.3
4	87.7	53.3	89.8	45.4
5	79.5	48.1	87.5	49.8
6	70.0	57.2	84.2	48.0
7	90.9	92.5	92.2	90.1
8	76.8	80.0	82.6	80.6
9	81.9	42.3	85.0	85.0
10	87.4	57.0	90.9	56.9
11	84.3	45.3	87.6	46.0
12	84.7	46.2	87.1	33.4
13	79.7	41.5	86.8	30.2
14	82.7	64.9	85.4	48.3
15	82.6	71.0	94.8	74.7
16	82.5	72.4	92.3	74.2
17	83.4	64.4	97.2	97.2
18	80.2	56.8	92.4	63.5
19	98.0	58.1	99.2	57.8
20	97.5	66.0	97.5	52.4
MEAN & SD	85.10 (5.90)	57.58 (13.71)	90.17 (4.63)	57.5 (19.80)



(a) Speaker purity as a function of number of speakers



(b) Cluster purity as a function of number of speakers

Figure 2. Fixed number of speakers ACP and ASP results.

In Figure 2 its noticeable that when using a fixed number of speakers, the speaker purity tend to decrease as there are more speakers and most results are below 60% and over 30%. Cluster purity with FLSD method doesn't seem to present a significant decrease as the number of speakers increase. The simple method appears to decrease slightly more than FLSD.

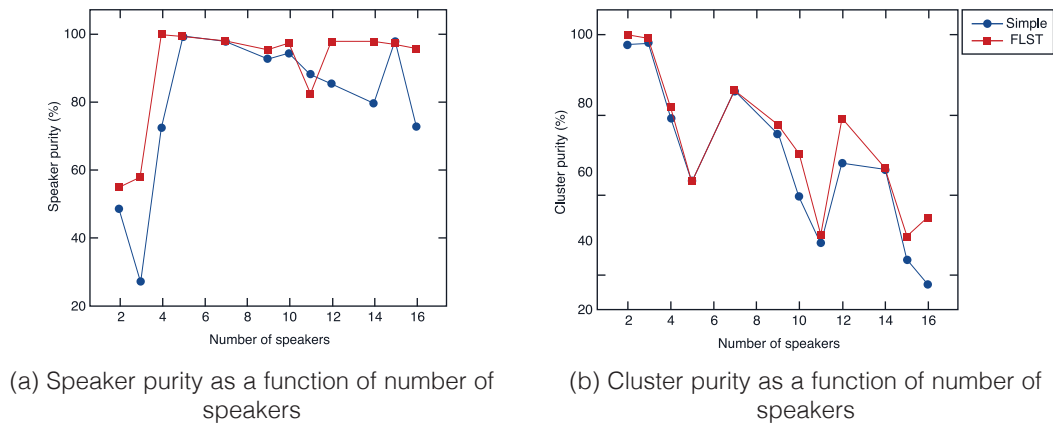


Figure 3. Speaker purity with a greater number of segments.

As shown in Figure 3, in both the base system and FLSD the speaker purity decreases with a greater number of segments. Cluster purity does not present a significant pattern of decrease with a greater number of segments. The base system and FLSD have similar behavior.

Conclusions

In this paper, we performed an analysis of the FLSD algorithm to experimentally validate its applicability in Costa Rican radio broadcast data, where the different conditions of the audio, and the presence of music segments and advertisements are challenging.

The FLSD diarization method had good results when there was a fixed number of speakers. On average, it matched the correct speaker 90.17% of the time. But the same method didn't perform similarly well with an unknown number of speakers as it only matched the speaker correctly 74.28% of the time. The outcomes show that with fewer participants, the result is better for both supervised and unsupervised diarization.

Cluster purity with a fixed number of speakers could have gotten worse due to the participants that appeared in ads for only short periods of time since the algorithm does not know beforehand how much participation a speaker had and it requires the speakers to appear more in order to identify them and associate them correctly. Also, some speakers had background music which makes it harder to correctly identify the speakers.

The speaker purity results are not as good as the cluster purity. It has to be taken into account that the database used is based on raw audio from a radio broadcast, which had music in different segments and signal losses due to the digital audio stream.

References

- [1] Barras, Claude, et al. "Multistage speaker diarization of broadcast news." *IEEE Transactions on Audio, Speech, and Language Processing* 14.5 (2006): 1505-1512.
- [2] Vavrek, Jozef, et al. "Classification of broadcast news audio data employing binary decision architecture." *Computing and Informatics* 36.4 (2017): 857-886.
- [3] García-Romero, Daniel, et al. "Speaker diarization using deep neural network embeddings." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.
- [4] Theodorou, Theodoros, Iosif Mporas, and Nikos Fakotakis. "An overview of automatic audio segmentation." *International Journal of Information Technology and Computer Science (IJITCS)* 6.11 (2014): 1.
- [5] Pleva, Matúš, and Jozef Juhár. "TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation." LREC. 2014.
- [6] Yilmaz, Emre, et al. "A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research." (2016).
- [7] Zgank, Andrej, Ana Zwitter Vitez, and Darinka Verdonik. "The Slovene BNSI Broadcast News database and reference speech corpus GOS: Towards the uniform guidelines for future work." LREC. 2014.
- [8] Nouza, Jan, Jindrich Zdansky, and Petr Cerva. "System for automatic collection, annotation and indexing of Czech broadcast speech with full-text search." *MELE – CON 2010-2010 15th IEEE Mediterranean Electrotechnical Conference*, 2010.
- [9] Federico, Marcello, Giordani, Dimitri and Coletti Paolo. "Development And Eval – uation Of An Italian Broadcast News Corpus." *European Language Resources Association (ELRA)*. 2000.
- [10] Giannakopoulos, Theodoros, and Sergios Petridis. "Fisher linear semi-discriminant analysis for speaker diarization." *IEEE transactions on audio, speech, and language processing* 20.7 (2012): 1913-1922.
- [11] Montazzolli, Sergio, Andre Adami, and Dante Barone. "An extension to Fisher Linear Semi-Discriminant analysis for Speaker Diarization." *2014 International Telecommunications Symposium (ITS)*. IEEE, 2014.
- [12] Sarafianos, Nikolaos, Theodoros Giannakopoulos, and Sergios Petridis. "Audiovisual speaker diarization using fisher linear semi-discriminant analysis." *Multimedia Tools and Applications* 75.1 (2016): 115-130.
- [13] Welling, Max. "Fisher linear discriminant analysis". Department of computer science, University of Toronto. Technical Report, 2005.
- [14] Giannakopoulos, Theodoros. "pyaudioanalysis: An open-source python library for audio signal analysis." *PloS one* 10.12 (2015): e0144610.

A low cost collision avoidance system based on a ToF camera for SLAM approaches

Un sistema de prevención de colisiones de bajo costo basado en una cámara ToF para enfoques SLAM

Dayron Romero-Godoy¹, David Sánchez-Rodríguez²,
Itziar Alonso-González³, Francisco Delgado-Rajó⁴

Romero-Godoy, D.; Sánchez-Rodríguez, D.; Alonso-González, I.; Delgado-Rajó, F. A low cost collision avoidance system based on a ToF camera for SLAM approaches. *Tecnología en marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 137-144.

 <https://doi.org/10.18845/tm.v35i8.6465>

- 1 Institute for Technological Development and Innovation in Communications. University of Las Palmas de Gran Canaria. Spain. E-mail: dayron.romero@idetic.eu
 <https://orcid.org/0000-0002-2805-2350>
- 2 Institute for Technological Development and Innovation in Communications. University of Las Palmas de Gran Canaria. Spain. Telematic Engineering Department. University of Las Palmas de Gran Canaria. Spain. E-mail: david.sanchez@ulpgc.es
 <https://orcid.org/0000-0003-2700-1591>
- 3 Institute for Technological Development and Innovation in Communications. University of Las Palmas de Gran Canaria. Spain. Telematic Engineering Department. University of Las Palmas de Gran Canaria. Spain. E-mail: itziar.alonso@ulpgc.es
 <https://orcid.org/0000-0001-8487-2559>
- 4 Institute for Technological Development and Innovation in Communications. University of Las Palmas de Gran Canaria. Spain. Telematic Engineering Department. University of Las Palmas de Gran Canaria. Spain. E-mail: paco.rajo@ulpgc.es
 <https://orcid.org/0000-0002-7262-7633>



Keywords

TOF camera; SBC; SLAM; obstacle detection; indoor localization.

Abstract

Indoor positioning is a problem that has not yet been solved efficiently and accurately. In outdoors the most effective solution is the Global Position System (GPS), but it cannot be used indoors due to the weakening of the signal, so other solutions have been studied. These approaches could be applied to define a map for the guidance of blind people, tourism or navigation for autonomous robots. In this paper, the study, design, implementation and evaluation of a robust obstacle detection and mapping system is proposed. Thus, it can be used to alert of near objects presence and avoid possible collisions in an indoor navigation. The system is based on a Time-of-Flight (ToF) camera and a Single Board Computer (SBC) like Raspberry PI or NVIDIA Jetson Nano. In order to evaluate the system several real experiments were carried out. This kind of system can be integrated on a wheelchair and help the handicapped person to move indoors or take data from an indoor environment and recreate it in a 2D or 3D images.

Palabras clave

Cámara TOF; SBC; SLAM; detección de obstáculos; localización en interiores.

Resumen

El posicionamiento en interiores es un problema que aún no se ha resuelto de manera eficiente y precisa. En exteriores la solución más eficaz es el Sistema de Posicionamiento Global (GPS), pero no se puede utilizar en interiores debido al debilitamiento de la señal, por lo que se han estudiado otras soluciones. Estos enfoques podrían aplicarse para definir un mapa para la orientación de personas ciegas, el turismo o la navegación para robots autónomos. En este trabajo se propone el estudio, diseño, implementación y evaluación de un sistema robusto de detección y mapeo de obstáculos. Por tanto, se puede utilizar para alertar de la presencia de objetos cercanos y evitar posibles colisiones en una navegación interior. El sistema se basa en una cámara de tiempo de vuelo (ToF) y una computadora de placa única (SBC) como Raspberry PI o NVIDIA Jetson Nano. Para evaluar el sistema se llevaron a cabo varios experimentos reales. Este tipo de sistema puede integrarse en una silla de ruedas y ayudar a la persona discapacitada a moverse en el interior o tomar datos de un entorno interior y recrearlos en imágenes 2D o 3D.

Introduction

In recent years, there has been a wide demand for services related to the location of people or objects. Currently, in outdoors, the Global Position System (GPS) is widely used and is capable of providing great accuracy when locating. Unfortunately, due to the attenuation of the satellite signal, the effectiveness of location in indoor environments is attenuated and then, less accurate, being even impossible to use to locate with certain accuracy. Therefore, in recent years

the development of indoor location systems has acquired great relevance using as a basis other types of sensors, such as the power level received from the signal of Wi-Fi access points, Bluetooth and information from inertial sensors. Today there are a variety of solutions, but none of them is optimal [2].

Thus, Simultaneous Localization And Mapping (SLAM) [11] is a technique that investigates the problem that raises the construction of mathematical, geo- metric or logical models of physical environments. For that, a mobile robot and a set of sensors and actuators are usually used to gather information about environment. As another way of saying this, SLAM seeks to solve the problems posed by placing a mobile robot in an unknown environment and position, and that the robot itself is capable of gradually building a consistent map of the environment while using this map to determine its own location. In general, SLAM is used to map an environment beforehand unknown by the automaton and at the same time estimate the path it is taking with the exclusive use of sensors it carries. There are several SLAM techniques such as using LIDAR [10], use of ultrasonic sensors or vSLAM (visual SLAM) [4]. Their applications can be focused to different fields, such as: robotics, tourism, rescue, navigation for blind people, etc. Smartphones and low-cost computers, such as Raspberry PI, have been recently considered as appropriate devices to easily obtain user information using various groups of embedded sensors. The vSLAM variant is based on a sensor camera to extract data from the environment. The cameras used can be either commercial cameras or Time of Flight (ToF) cameras [5, 3].

A ToF camera is capable of providing a depth-sensing of a scene which each gathered pixel stores its X, Y, and Z coordinates in the image. Z value represents the distance from the camera to the point of focus [6]. This type of camera provides high quality measurements, and it is ideal for applications that requires high performance and high accuracy. With this system, stable measurements in both accuracy and recurrence are achieved, even with objects of different colors and reflectivity within the image.

On the other hand, point cloud processing is an important aspect of many systems implemented in real world. As such, a wide variety of point-based approaches have been proposed and reporting steady benchmark improvements over time [1]. Most 3D scanners give raw scanned data in form of point cloud format. The point cloud produced by 3D scanners are visualized for the ease of measurement or even representation. A point cloud is a set of data points in a 3D coordinate system, consisting by x, y, and z coordinates values. They are used to represent the surface of an object, hence, do not contain data of any internal features like color or materials. With this kind of data is possible to develop algorithms that use point clouds for certain purposes.

The aim of this work is to developed a low-cost system composed by a ToF camera and a SBC (single Board Computer) to alert of near objects presence and avoid possible collisions in an indoor navigation. In addition, the scene can be recreated from the cloud point processing.

This work is structured as follows: section 2 describes the used methodology to alert about possible collisions. In section 3, the system architecture is proposed. After that, in section 4, the results of experiments are shown and discussed. Finally, section 5 contains the conclusion and future work.

Methodology

In this section the process and concepts needed to achieve point cloud processing is explained. In addition, the detection of possible collides and representation of point clouds in 3D images is described. The methodology followed in this work is shown in figure 1, which is developed in three phases: point cloud extraction, filtering data, and collision detection and 3D representation.

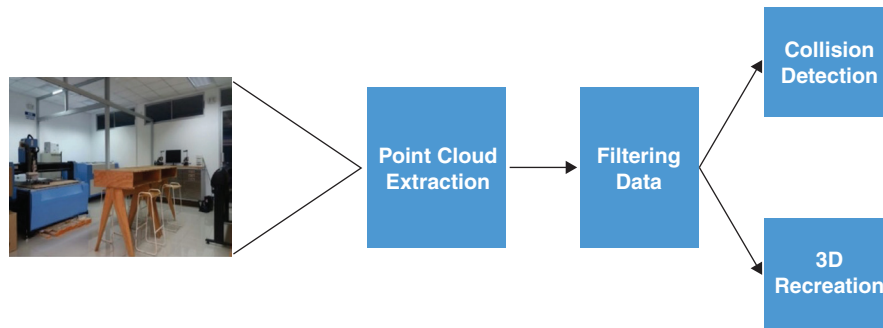


Figure 1. The proposed methodology.

The cloud point extraction phase is carried out using the API manufacturer's API. With this library, the camera is activated and the data collection is carried out. In addition, a Python library named NumPy is used for working with arrays because it provides optimized numerical computations, and therefore, it is faster than traditional operations. On the other hand, sometimes the ToF camera can return outliers which are pixel information that not match with the scene. A pixel is considered as outlier if the absolute value of the Z-value mean of their nearest neighbours is five times bigger than the Z-value of pixel. Therefore, these pixels are removed from point cloud in the filtering data phase. Next, and in order to carry out the collision detection, a threshold can be set which represents the minimum distance to avoid a collision. Using the NumPy library, the difference among each point and the threshold is computed. If it is equal or less than zero a possible collision can occur, and therefore, an alert message is indicated. If the threshold is not set, the system return the distance to closer object. Due to the camera specifications, the minimum threshold that can be set is 0.1 m. Lastly, an algorithm based on the PointCloud Library [7] is used to visualize a 3D scene representation.

System Architecture

In this section the system architecture is proposed, figure 2. Basically, the system is formed by a ToF camera and a SBC. In addition, a server has been used to store the gathered point clouds and visualize the 3D representation. Next, both ToF camera and SBC are described.



Figure 2. System architecture.

ToF Camera

As ToF camera, a CamBoard pico flexx [6] was used. It is a peripheral depth sensor with USB connection that can be integrated into a mobile device. The characteristics of this camera are a configurable capture rate that goes from 5fps up to 45fps, a measurement range up to 4 meters using 5fps, and a resolution of 224 x 171px. It has support in Android, MacOS, Windows and Linux operating systems.

Single Board Computer

Two SBC were used to test the proposed methodology: a NVIDIA Jetson Nano [8] and a Raspberry Pi [9].

The NVIDIA Jetson Nano is a SBC designed for the development of artificial intelligence applications, ideal for robotics, image processing, object detection, segmentation and many more applications. It uses a Quad Core ARM Cortex- A57 which is 1.43 GHz powerful 64 bit quad core processor, and 128 CUDA core GPU. As for operating systems, it can run Ubuntu and other Linux operating systems. The operating system provided by NVIDIA for the Jetson Nano is adapted from Ubuntu.

On the other hand, Raspberry Pi 4 model B 4G is a SBC developed in the United Kingdom by the Raspberry Pi Foundation, with the aim of promoting computer education in schools. It uses a quad-core Arm Cortex-A53 which uses 1.5 GHz 64 bit quad core processor. The software is open source, with its Raspbian operating system, a version adapted from Debian for this type of SBC.

Results and Discussion

In this section, the methodology and the system architecture were evaluated through several experiments.

Collision Detection

In order to evaluate the collision detection several experiments were carried out varying the distance among the system and a person located in the scene. When the test begins the person starts moving from right to the center of scene and return to the original place. Figure 3 shows the distance to the person while is moving. Both Nvidia Jetson Nano and Raspberry Pi 4 have been used as SBC. Both platforms achieves similar results, nevertheless fewer point clouds are processed using the Raspberry PI because it has less computational power. Therefore, Nvidia Jetson Nano is better for real-time applications. On the other hand, in order to evaluate the accuracy of the system, a object was located to 1, 1.2, 1.5, 1.7 and 2.0 meters from ToF camera. For each position 100 point clouds were processed. Table 1 shows the average results obtained by the system. As can be seen, the system has a millimeters accuracy.

Table 1. Real distance vs estimated distance to object.

Real distance (m)	Averaged estimated distance (m)
1.00	1.002 ± 0.005
1.20	1.203 ± 0.005
1.50	1.505 ± 0.005
1.70	1.703 ± 0.005
2	2.004 ± 0.005

3D Scene Representation

This section will show a scene consisting of a bed with an electric guitar on the left and a Spanish guitar on the right. It is important to note that the electric guitar is somewhat more advanced than the Spanish one. The aim of this algorithm is to recreate this scene in detail by means of a point cloud using PointCloud Library. It should be noted that with the data collected through a single frame is possible to change and rotate the perspective of the point cloud. Figure 4 and 5 show the real scene taken from ToF camera, and its representation from the point cloud. As can be seen, the perspective of the generated point cloud has been rotated to appreciate the scene from another perspective, and thus differentiate the distance between both guitars. It is worth noting that the cavity of the Spanish guitar's soundboard can also be observed.

```

The possible collision is at 1.1141916513442993 meters
The collision is at your right
2020-11-02 15:57:06.501186 : There is a possible collision ahead
The possible collision is at 1.0462698936462402 meters
The collision is at your right
2020-11-02 15:57:06.696506 : There is a possible collision ahead
The possible collision is at 1.0947818756103516 meters
The collision is at your right
2020-11-02 15:57:06.894520 : There is a possible collision ahead
The possible collision is at 1.1014058589935303 meters
The collision is at your right
2020-11-02 15:57:07.093625 : There is a possible collision ahead
The possible collision is at 1.1202033758163452 meters
The collision is at your right
2020-11-02 15:57:07.312064 : There is a possible collision ahead
The possible collision is at 1.0761597156524658 meters
The collision is in front of you
2020-11-02 15:57:07.491666 : There is a possible collision ahead
The possible collision is at 1.0344611406326294 meters
The collision is at your right
2020-11-02 15:57:07.696132 : There is a possible collision ahead
The possible collision is at 1.0751615762710571 meters
The collision is at your right
2020-11-02 15:57:07.894314 : There is a possible collision ahead
The possible collision is at 1.083804726600647 meters
The collision is at your right
2020-11-02 15:57:08.100380 : There is a possible collision ahead
The possible collision is at 0.9497186541557312 meters
The collision is at your right

```

Figure 3. Obstacle detection in real-time.

Conclusion and Future Work

In this paper a methodology based on a ToF camera and a SBC is proposed to announce or alert about possible collisions in indoor environments. The approach provides a high accuracy, and besides a 3D representation of the scene can be generated from a point cloud. Therefore, it could be added to a wheelchair to provide assistance to people with functional diversity in indoor environments. Moreover, two low cost SBC were evaluated delivering better results the Nvidia Jetson Nano due to its high processing speed and better components.

In our ongoing work, we are planning to generate a map of the environment from several point clouds using the IndoorGML standard. In addition, OpenCV library can be used for the identification of both people and objects, and therefore a better feedback about obstacle is notified to the user.



Figure 4. Real scene.

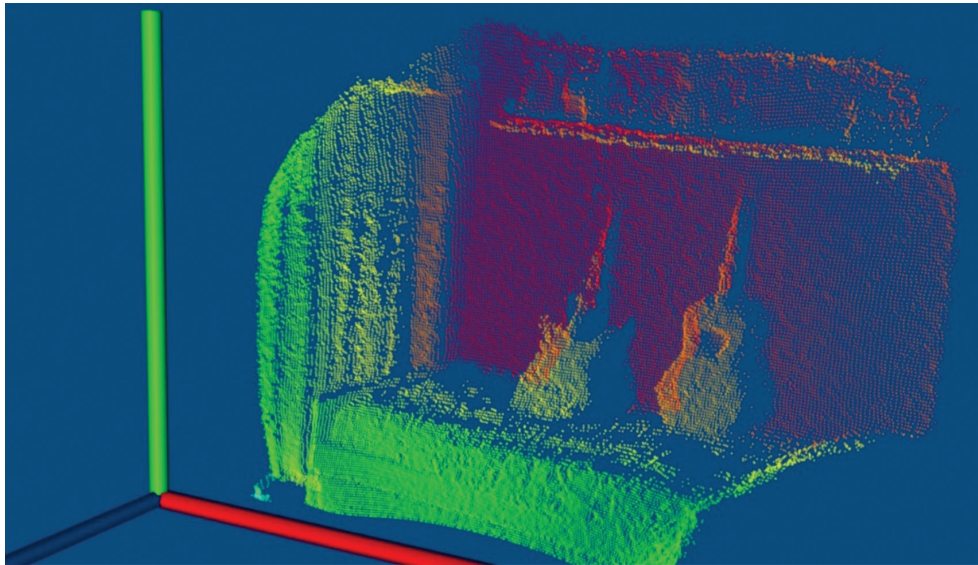


Figure 5. Scene recreation from a point cloud.

Acknowledgments

This research has been partially funded by the Consejería de Economía, Conocimiento y Empleo del Gobierno de Canarias, Agencia Canaria de Investigación, Innovación y Sociedad de la Información under projects ProID2020010009 and CEI2020-08, Spain.

References

- [1] Bello, S.A., Yu, S., Wang, C., Adam, J.M., Li, J.: deep learning on 3d point clouds. *Remote Sensing* 12(11), 1729 (2020)
- [2] Gao, C., Harle, R.: Semi-automated signal surveying using smartphones and floor- plans. *IEEE Transactions on Mobile Computing* 17(8), 1952–1965 (2017)
- [3] He, Y., Chen, S.: Recent advances in 3d data acquisition and processing by time-of-flight camera. *IEEE Access* 7, 12495–12510 (2019). <https://doi.org/10.1109/ACCESS.2019.2891693>
- [4] Karlsson, N., Di Bernardo, E., Ostrowski, J., Goncalves, L., Pirjanian, P., Munich, M.E.: The vslam algorithm for robust localization and mapping. In: *Proceedings of the 2005 IEEE international conference on robotics and automation*. pp. 24–29. IEEE (2005)
- [5] Paredes, J.A., A'lvarez, F.J., Aguilera, T., Villadangos, J.M.: 3d indoor positioning of uavs with spread spectrum ultrasound and time-of-flight cameras. *Sensors* 18(1), 89 (2018)
- [6] Pasinetti, S., Hassan, M.M., Eberhardt, J., Lancini, M., Docchio, F., Sansoni, G.: Performance analysis of the pmd camboard picoflexx time-of-flight camera for markerless motion capture applications. *IEEE Transactions on Instrumentation and Measurement* 68(11), 4456–4471 (2019)
- [7] Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: *2011 IEEE international conference on robotics and automation*. pp. 1–4. IEEE (2011)
- [8] Tao Peng, Dingnan Zhang, D.L.N.H.J.L.: An evaluation of embedded gpu systems for visual slam algorithms. *Electronic Imaging* p. 325 (2020). <https://doi.org/10.2352/ISSN.2470-1173.2020.6.IRIACV-325>
- [9] Vujović, V., Maksimović, M.: Raspberry pi as a wireless sensor node: Performances and constraints. In: *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. pp. 1013–1018 (2014). <https://doi.org/10.1109/MIPRO.2014.6859717>
- [10] Wang, Y.T., Peng, C.C., Ravankar, A.A., Ravankar, A.: A single lidar-based feature fusion indoor localization algorithm. *Sensors* 18(4), 1294 (2018)
- [11] Zhang, H., Ye, C.: An indoor wayfinding system based on geometric features aided graph slam for the visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25(9), 1592–1604 (2017). <https://doi.org/10.1109/TNSRE.2017.2682265>

A first study on age classification of costa rican speakers based on acoustic vowel analysis

Un primer estudio sobre clasificación por
edades de hablantes de costarricense
basado en análisis de vocales acústicas

Victor Yeom-Song¹, Marvin Coto-Jiménez²

Yeom-Song, V.; Coto-Jiménez, M. A first study on age classification of costa rican speakers based on acoustic vowel analysis. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 145-152.

 <https://doi.org/10.18845/tm.v35i8.6466>

- 1 University of Costa Rica. Costa Rica. E-mail: victor.yeom@ucr.ac.cr
 <https://orcid.org/0000-0003-4172-1536>
- 2 University of Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr
 <https://orcid.org/0000-0002-6833-9938>

Keywords

Age recognition; children's speech; classification; vowel analysis.

Abstract

According to several studies, children's speech is more dynamic and inconsistent compared to an adult's speech. This aspect can be considered in the task of recognizing the age of the person who speaks and of great importance in many applications, such as human-computer interaction, security on Internet and education assistants. Those applications have a dependency on language and accent, due to the different sounds and styles that characterize the speakers. This paper presents the initial results on the identification of Costa Rican children's speech, in a database created for this purpose, consisting of words pronounced by adults and children of several ages. For this first study we chose the most common vowel of the language, and extract a set of common acoustic features to determine its applicability in distinguishing between adults and children of an age range. The outcome results shows promising results in the classification using a single vowel, that improves according to the number of vowels used to extract the acoustic features. This means that an automatic system could be able to improve its capacity to identify age as more speech information is received and transcribed, but cannot be very accurate in short interactions.

Palabras clave

Reconocimiento de edad; habla infantil; clasificación; análisis de vocales.

Resumen

Según varios estudios, el habla de los niños es más dinámica e inconsistente en comparación con el habla de un adulto. Este aspecto se puede considerar en la tarea de reconocer la edad de la persona que habla y de gran importancia en muchas aplicaciones, como la interacción humano-computadora, la seguridad en Internet y los asistentes educativos. Esas aplicaciones tienen una dependencia del lenguaje y el acento, debido a los diferentes sonidos y estilos que caracterizan a los hablantes. Este trabajo presenta los resultados iniciales sobre la identificación del habla infantil costarricense, en una base de datos creada para tal fin, que consta de palabras pronunciadas por adultos y niños de distintas edades. Para este primer estudio, elegimos la vocal más común del idioma y extraemos un conjunto de características acústicas comunes para determinar su aplicabilidad para distinguir entre adultos y niños de un rango de edad. Los resultados obtenidos muestran resultados prometedores en la clasificación utilizando una sola vocal, que mejora según el número de vocales utilizadas para extraer las características acústicas. Esto significa que un sistema automático podría mejorar su capacidad para identificar la edad a medida que se recibe y transcribe más información del habla, pero no puede ser muy preciso en interacciones breves.

Introduction

Speech signals contain information of several kinds. The most visible part is the linguistic content but, additionally, there exists paralinguistic information associated with the message, such as the speaker's accent, gender, age, or emotional state [1]. Even the state of health, including those ones related to the recent COVID-19 outbreak, have been successfully investigated using speech signals [2] or sound associated with the vocal tract [3].

Research in the processing of such paralinguistic information has grown considerably over the last two decades, including, more recently, children's speech [4]. One particular task of this processing is the age and gender recognition from speech recordings.

Automatic recognition of the paralinguistic information that allows the identification of children from their speech can be of benefit in many application areas. For example, to guide a child-computer (or child-robot) interaction, automatically adapt content, enhance child security in communications and inter- actions through the Internet, and educational applications [4].

Most current Automatic Speech Recognition (ASR) systems are not particularly accurate with children's speech, especially those in preschool age [5]. The main explanation for this issue is the acoustic mismatch between children's speech and the information used to train the recognizers. To overcome this problem it is vital to analyze and describe the characteristics of children's speech in each language, and develop systems that can adapt in terms of recognizing and, in the case of two way communication, even adapt its vocabulary when a child is speaking.

As children grow, these automatic recognition rate improve, as well as the general understanding of the words for native speakers [6], which can be related to the closer characteristics of the sounds and general articulation during the years of development.

Those characteristics of children's speech vary rapidly as a function of age, due to the anatomical and physiological changes occurring during their development [7]. Although a quantitative study of a cross lingual validation of this statement has not been addressed, it is clear that the common changes during a child's development affects and improves the acoustical characteristics of their speech.

Acoustic analysis of children speech, especially for younger ages, is a challenging task from the very beginning of data recording, where an interaction strategy should be applied to establish a proper environment for the participation of children. But, in terms of developing human-computer interaction for this population, including speech recognition, it is mandatory to establish ways of automatic identification of children using their voices' acoustic characteristics. For this reason, in this work we conducted a first study on Costa Rican speakers' age classification, particularly in two classes: children and adults. Our proposal seeks to explore the possibility of performing this classification using acoustic characteristics of the most common vowel in the Spanish language.

Related work

The analysis of formants, pitch and duration in the English language can be tracked back [8] for children of 5 to 18 years old. This study also pioneered the analysis of vowel characteristics for this population in the English language, combined with duration, and spectral variations.

The study of vowels in several languages in terms of recognition of sounds has been presented in [9] with a recognition rate of 76.25%. In the Russian language, the vowels pronounced by children of 6 and 7 years have been analyzed in [6]. The main differences between the vowels of this population and Russian adults are longer duration and high pitches. Also, the formant structure is not formed completely in children of these age ranges.

For the case of the English language, a characterization of children's vowels in terms of variability in repetitions has been presented in [7], with higher variability in the younger population. As this study shows, the majority of the analysis of children's speech dealt with vowel duration, pitch, and formants. The analysis of consonants is less common in the literature. Other acoustic characteristics, such as co-articulation, have also been studied in [10].

Also for the English language, in [5], acoustic characteristics of young children's speech were studied as a function of age. The fundamental frequency, formants and vowel duration for vowels were found to show age-dependent trends; particularly the variability of those parameters, with less significant changes in pitch.

For the Italian language, changes in acoustic characteristics of children have also been presented in [11], confirming that characteristics of children's speech change with age, and that spectral and temporal variability decrease as age increases. The variability shows substantial differences in children at an early age compared to children at a late age.

In terms of classification of children and adult speech, recent works have made use of pitch and formants with spectral coefficients [12]. The accuracy of the classification was above 97%. In the Italian language, recent experiences have reported accuracy higher than 80% [13] in children and adult recognition.

For the Portuguese language, a classifier comparison for children and adult speech recognition using Perceptual Linear Prediction coefficients and pitch as features were presented in [14]. Using a small amount of training data the accuracy of the best classifiers were as high as 97.4%.

In our work, we extend the application of acoustic analysis of vowels commonly performed in other languages for Costa Rican children, and compare some classifiers' performance to detect the voice of an adult or a child using only the acoustic analysis of such sounds.

The rest of this paper is organized as follows: Section 2 presents the experimental setup of our work. Section 3 shows the Results and Discussion, and finally in Section 4 we present the conclusions and future work.

Experimental Setup

Database

Two sessions for recording isolated words of children with ages between 4 to 12 years old were conducted, with a strategy of interaction with pictograms. These pictograms included simple words according to the development of the language of children in the first years. The same set of words was repeated for all the children, and a group of 8 adults. All the participants are native Costa Rican Spanish speakers.

The children were divided into two groups: Children in early childhood (those between within ages 4 to 7 years old), and Children in late childhood (those between ages 8 to 12 years old).

The selection of the same words of all the participants was manually edited and then segmented using the Praat system [15]. In this system, the temporal marks for each sound and its corresponding features can be extracted. For our experiments, we began with the selection of information of the five vowels of Spanish language: /a/, /e/, /i/, /o/, /u/.

To build the database for the experimentation with classifiers, we selected the most common vowel of the Spanish language: /e/ [16]. Then, we extracted subsets of three, five and ten randomly selected vowels within each age group of one single speaker each time. With this procedure, we pretend to emulate the vocal emission of any word or phrase, where it is not possible to establish prior information on the upcoming vowels and its characteristics.

Features

For each vowel, we extract the most common acoustic analysis, according to previous references, from sets of 3, 5 and 10 vowels: Average fundamental frequency (f_0) in Hz, Minimum f_0 in Hz, Maximum f_0 in Hz, Average duration in seconds, Minimum duration in seconds, Maximum

duration in seconds, Average position of the first formant of the spectrum (F1) in Hz, Minimum and Maximum of F1 in Hz, and the same measures of the second and third formants (F2 and F3). Detailed description of the algorithms applied in the features extraction can be found in [17].

Results and Discussion

In this section, we present the results obtained for the classification task with three different classifiers: Random Forest, Support Vector Machines (SVM), and k-Nearest Neighbors (kNN). It is to be noted that for kNN, we used $k=1$, so it is just a Nearest Neighbor classifier. We can see the classification results using 3 vowels in Table 1. We can see that, in this case, the Random Forest has the best classification performance overall, even though the scores, in general, are really similar for each classification task according to the age ranges. The best case out of all of the ones presented belongs to the Random Forest classifier classifying children in early childhood and adults, which is to be expected since the acoustic features under study present the most variability within these age groups.

The classification results from using 5 vowels are shown in Table 2. There is an evident increase in all the metrics across the board, but it should be noted that the Random Forest classifier for early children and adults had the least amount of improvement. The best classification case in general still belongs to the Random Forest for classification between children in late childhood and adults, even as the SVM has the best performance in the two other cases. This outcome could just be a result of the random seed used for the Random Forest, and may not be entirely indicative of Random Forest being completely better than the other classifiers.

It should be noted that with five vowels, each classifier's performance is really similar for each respective case, with pretty much less than a 2% variation in accuracy for most cases. Thus, it really could not be said that one classifier is definitely better than another one in this task.

Table 1. Classification using data from three /e/ vowels.

Random Forest				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	90.00	0.90	0.90	0.90
Late childhood	85.00	0.85	0.85	0.85
All	87.10	0.87	0.87	0.87
SVM				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	86.67	0.87	0.87	0.87
Late childhood	83.89	0.85	0.84	0.84
All	87.10	0.87	0.87	0.87
kNN				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	85.00	0.85	0.85	0.85
Late childhood	83.33	0.84	0.83	0.83
All	86.13	0.87	0.86	0.86

Table 2. Classification using data from five /e/ vowels.

Random Forest				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	91.67	0.92	0.92	0.92
Late childhood	94.97	0.95	0.95	0.95
All	91.29	0.91	0.91	0.91
SVM				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	92.50	0.93	0.93	0.93
Late childhood	91.06	0.91	0.91	0.91
All	92.26	0.92	0.92	0.92
kNN				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	90.83	0.91	0.91	0.91
Late childhood	92.74	0.93	0.93	0.93
All	90.00	0.90	0.90	0.90

The classification results using 10 vowels are shown in Table 3. We can see that, in general, the results are even better than with 5 vowels, and that now the best classification performance belongs to the SVM in classification between early children and adults, with an astounding 99% accuracy. In other cases, other classifiers have better performance. Again, the difference in the results between one case and another differs by less than 2%, so there would not be one clear winner for classification.

Table 3. Classification using data from ten /e/ vowels.

Random Forest				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	98.33	0.98	0.98	0.98
Late childhood	97.22	0.97	0.97	0.97
All	97.74	0.98	0.98	0.98
SVM				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	99.17	0.99	0.99	0.99
Late childhood	96.11	0.96	0.96	0.96
All	95.81	0.96	0.96	0.96
kNN				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	98.33	0.98	0.98	0.98
Late childhood	98.33	0.98	0.98	0.98
All	97.10	0.97	0.97	0.97

One aspect to be noted with the results is that in each case, all the metrics for accuracy, precision, and recall (and since the latter 2 are equal, F-measure as well) are numerically pretty much the same. This tells us that in each case the rate of false positives is equal to the rate of false negatives, which, in turn, means that the classification task is equally prone to type I and type II errors.

With the presented results, it can be seen that the classification of speech for children and adults is a very manageable task. In fact, with 5 and 10 vowels, the classifiers produced similar results between one another, with less than 2% variation in the used metrics. Since three different classifiers with reasonably different architectures managed to produce such similar results, it could be an indicator of possible ease in the classification stage.

As expected, the better classification results are obtained with 10 vowels as summarized in Figure 1, but the results with 3 and 5 vowels are really good as well. Since waiting for 10 utterances of a single vowel could be considered slow for real-life use, one could look at the previous results and consider using them depending on the application.

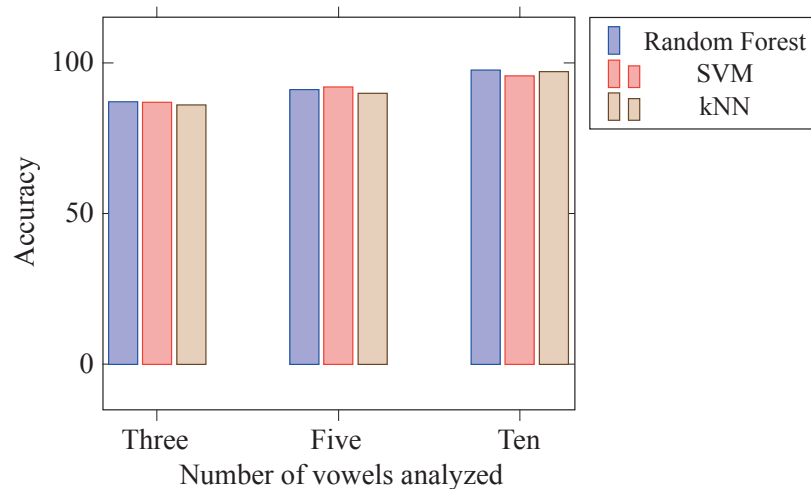


Figure 1. Results for each classifier for children in all age range and the amount of vowels analyzed.

Conclusions

The classification of speech between children and adults is a task with many potential applications, ranging from recommender systems to web security. As such, a study on the nature of the task itself is useful to explore its viability in implementation. This work offers a first approach to such a type of research for Costa Rican Spanish, using three different classifiers to assess the possibility of the task within a framework of acoustic features.

Using only the vowel /e/ (the most frequent in the Spanish language), the results obtained are promising, considering that with only 3, 5, or 10 vowels produced a considerably high performance across the experiments, with over 80% accuracy using only 3 vowels. In general, the best results were obtained with 10 vowels. However, this doesn't necessarily mean that for real-life use, this is the best way to classify voice, as it could be considered slow depending on the application.



As for the task itself, the use of three different classifier types shows that it is quite manageable since all the classifiers produced really similar results in each study case. Thus, the implementation of a real-time speech classifier wouldn't be too far-fetched with current technology. Further exploration of the nature of the data could show better results in the future.

For future work, it is possible to consider the use of combinations of vowels for the task of classification, to study better classification boundaries or if.

Additionally, the application of other classifiers or set of parameters to achieve better results and contemplating the possibility of a real time age recognition system for Costa Rican Spanish.

References

- [1] Safavi, Saeid, Martin Russell, and Peter Jančovič. "Automatic speaker, age-group and gender identification from children's speech." *Computer Speech & Language* 50 (2018): 141-156.
- [2] Schuller, Bjorn W., et al. "Covid-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 Corona crisis." *arXiv preprint arXiv:2003.11117* (2020).
- [3] Imran, Ali, et al. "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app." *Informatics in Medicine Unlocked* (2020)
- [4] Safavi, Saeid, et al. "Identification of gender from children's speech by computers and humans." *INTERSPEECH*. 2013.
- [5] Yildirim, Serdar, et al. "Acoustic analysis of preschool children's speech." *Proc.15th ICPHS*. 2003.
- [6] Lyakso, Elena E., Olga V. Frolova, and Aleks S. Grigoriev. "The acoustic characteristics of Russian vowels in children of 6 and 7 years of age." *Tenth Annual Conference of the International Speech Communication Association*. 2009.
- [7] Gerosa, Matteo, et al. "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition." *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 1. IEEE, 2006.
- [8] Lee, Sungbok, Alexandros Potamianos, and Shrikanth Narayanan. "Analysis of children's speech: Duration, pitch and formants." *Fifth European Conference on Speech Communication and Technology*. 1997.
- [9] Ting, Hua Nong, and Jasmy Yunus. "Speaker-independent Malay vowel recognition of children using multi-layer perceptron." *2004 IEEE Region 10 Conference TENCN 2004*. IEEE, 2004.
- [10] Katz, William F., and Sneha Bharadwaj. "Coarticulation in fricative-vowel syllables produced by children and adults: A preliminary report." *Clinical linguistics & phonetics* 15.1-2 (2001): 139-143.
- [11] Gerosa, Matteo, Diego Giuliani, and Fabio Brugnara. "Acoustic variability and automatic recognition of children's speech." *Speech Communication* 49.10-11 (2007): 847-860.
- [12] Zeng, Yumin, and Yi Zhang. "Robust children and adults speech classification." *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*. Vol. 4. IEEE, 2007.
- [13] Massarente, Enrico. "Classificazione automatica della voce in ambito logopedico: training e testing di un algoritmo per discriminare la voce adulta da quella dei bambini." (2015).
- [14] Martins, Rui, et al. "Detection of Children's Voices." *I Iberian SLTech 2009*: 77.
- [15] Goldman, Jean-Philippe. "EasyAlign: an automatic phonetic alignment tool under Praat." *Interspeech'11, 12th Annual Conference of the International Speech Communication Association*. 2011.
- [16] Guirao, Miguelina, and María García Jurado. "Frequency of occurrence of phonemes in American Spanish." *Revue quebecoise de linguistique* 19.2 (1990): 135-149.
- [17] Boersma, P. Weenink, D. "Praat: doing phonetics by computer" [Computer program]. Version 6.0.37, retrieved May 2020 from <http://www.praat.org/>.

An experimental study on footsteps sound recognition as biometric under noisy conditions

Un estudio experimental sobre el reconocimiento del sonido de las pisadas como biométrico en condiciones ruidosas

Marisol Zeledón-Córdoba¹, Carolina Paniagua-Peñaranda², Marvin Coto-Jiménez³

Zeledón-Córdoba, M.; Paniagua-Peñaranda, C.; Coto-Jiménez, M. An experimental study on footsteps sound recognition as biometric under noisy conditions. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 153-161.

 <https://doi.org/10.18845/tm.v35i8.6467>

1 Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: marisol.zeledon@ucr.ac.cr

 <https://orcid.org/0000-0001-7481-0207>

2 Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: carolina.paniaguapenaranda@ucr.ac.cr

3 Electrical Engineering Department. University of Costa Rica. Costa Rica.
E-mail: marvin.coto@ucr.ac.cr

 <https://orcid.org/0000-0002-6833-9938>

Keywords

Biometric; classification; footsteps; sounds.

Abstract

The experimentation of footsteps as a biometric has a short history of about two decades. The process of identification of a person is based on the study of footstep signals captured when walking over a sensing area, and the registering of sounds, pressure, vibration, or a combination of these measures. Application of this biometric can emerge in security systems, that identify persons who enter or leave a space, and in providing help to elderly and disabled persons. In this paper, we are focused in the exploration of pure audio signals of footsteps and the robustness of a person's classification under noisy conditions. We present a comparison between four well-known classifiers and three kinds of noise, applied at different signal to noise ratio. Results are reported in terms of accuracy in the detection an users, showing different levels of sensibility according to the kind and level of noise.

Palabras clave

Biométrico; clasificación; pisadas; sonidos.

Resumen

La experimentación de las pisadas como biométrico tiene una breve historia de unas dos décadas. El proceso de identificación de una persona se basa en el estudio de las señales de pisadas capturadas al caminar sobre un área de detección y el registro de sonidos, presión, vibración o una combinación de estas medidas. La aplicación de esta biometría puede surgir en los sistemas de seguridad, que identifican a las personas que entran o salen de un espacio, y en la prestación de ayuda a las personas mayores y discapacitadas. En este artículo, nos centramos en la exploración de señales de audio puras de pasos y la solidez de la clasificación de una persona en condiciones ruidosas. Presentamos una comparación entre cuatro clasificadores conocidos y tres tipos de ruido, aplicados a diferentes relaciones señal / ruido. Los resultados se informan en términos de precisión en la detección de los usuarios, mostrando diferentes niveles de sensibilidad según el tipo y nivel de ruido.

Introduction

In the area of biometrics, the most common elements for verifying a person's identity are fingerprints and face, usually applied in smartphones. Others, such as iris identification have also been successfully used in applications in airports [1]. These are examples of the physiological group of biometrics. Apart from the physiological group, the use of behavioral biometrics, such as voice recognition has also received considerable attention in recent years. Footsteps signals are an- other example of behavioral biometrics, with a shorter history of active research. Footstep signals are signals collected from people walking over an instrumented sensing area [1], with the aim of classification or analysis of the individuals. The proposal for analyzing and evaluating such signals was introduced by [2], remarking its simplicity for clinical practice.

But it was until 1997 that the first experiments with sensors in an active floor were presented [3]. Since then, several researchers have offered different approaches to the application of footsteps or gait analysis as a biometric. These works have demonstrated the real potential of the footstep biometric [4].

The human identification and analysis of footsteps can have applications in medicine, surveillance, sport shoe industry, smart homes, and multimedia [5, 6]. Given the relatively recent experiences with this measure, many recent studies have presented the building of datasets with particular measurements of sensors, features, classifiers, and conditions [5].

Additionally, there are areas of concern in their usage in terms of practicality, privacy, and security [7]. For this reason, the scientific literature on footsteps and gait analysis has grown significantly in the past years, with sensors based on vision, sound, pressure, and accelerometry [8], with the corresponding set of features.

In this work, we present the building of a dataset of footsteps sounds and report on the first experiments on the robustness of several classifiers for this biometric in the presence of noise. For this experience, we consider only the discrimination of footsteps between two individuals of similar age and gender, where the challenges are more relevant.

Related work

The assessment of signals for their application as a biometric involves developing datasets with a large number of labelled examples [9]. For lesser researched biometrics, such as footsteps, new databases are needed in order to assess the accuracy and study other practical aspects. In particular, the registering of only the sound of footsteps in a distant microphone requires the development of such datasets.

Several studies have reported the potential of using gait information to distinguish between people. Previous references have reported its use in criminal cases to identify perpetrators based on their walking behavior, and also in the identification of patterns such as the Parkinson [8].

Previous reports have achieved around 80% to 90% [6, 1] in the identification of individuals using footstep information. But a wide variety of conditions do not allow a precise comparison of the efforts or the establishment of benchmarks in this field for unexplored sensing conditions, such as the pure sound of the footsteps.

Among the many conditions that can affect the performance of the sound of footsteps as a biometric are the different types of footwear worn, like heels, sneakers, leathers or even barefooted, and the corresponding sound in different grounds of concrete, wood or other materials [10, 11].

A comparison of features and sensors during the first decade of studies in footsteps recognition, as presented in [5] has not registered the application of pure sound signals in the identification of individuals. For this reason, the exploration of features derived from sound signals requires the building of a dataset to explore it. With microphones being one of the simplest and cheapest sensors, its application in biometric of footsteps can be particularly useful.

One of the most common issues related to sound signals is the presence of noise. In a real-life application of biometrics with footsteps sounds, the presence of noise and sounds other than footsteps is continuous. In this work, we explore the building and testing of the sound of footsteps as a biometric, using a single microphone to register the distant sound of footsteps, and compare the performance of several classifiers under noisy conditions. Being the first dataset and the focus on the impact of noise, we consider the simplest case of binary classification of two users.

The rest of this paper is organized as follows: Section 2 presents the experimental setup for building the dataset and the experimentation. Section 3 presents the results and discussion, and finally Section 3 presents the conclusions and future work.

Experimental Setup

For evaluating how background noise can affect the identification of an individual using footsteps sounds, we developed recording sessions with several male and female participants. For this work, we only considered pairs of recordings of two women's footsteps, where the distinction between two people is more challenging, according to our first experiences.

The recordings were made using a single microphone with an Omnidirectional pattern, and the participants were asked to walk naturally and continually in a circle of 1.5 m around the microphone. Figure 1 shows the basic setup for the sessions. Each participant recorded about fifteen minutes of her footsteps.



Figure 1. Recording session.

Each recording was then edited in segments of five seconds, in order to capture at least 3 footstep sounds in each segment. For each type of noise considered in this work, specific SNR levels were added to each file. This means that there are several versions of the sound for each audio segment: those with specific SNR level of each type of noise, and the one without any noise added (clean version). For each segment of the whole set of conditions, a set of features were extracted, corresponding to several categories of audio descriptors. For example, the energy, zero crossing rate, entropy, MFCC and chroma features. The complete description of the features corresponds to those presented in [12].

For each participant, about 3000 files were generated, considering the whole set of conditions. The set of features for each of such files were tested using four common classifiers: K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and Support Vector Machine (SVM). In each case, a ten-fold cross-validation was used to assess the accuracy of each classifier under each condition.

Results and Discussion

This section presents the results obtained from the different classifiers and the noise levels analyzed. In each of the tables, a noise level with the accuracy percentage obtained for that level is presented. The results are reported in terms of the distinction between two people. In other words, all the results correspond to binary classification as a way to analyze the first case of an experimental study that uses only audio signals for biometric identification.

Table 1. Accuracy of classifiers for different levels of Babble Noise.

Clean				
	KNN	N. Bayes	R. Forest	SVM
P_1 vrs P_2	96.6%	96.4%	98.2%	98.1%
P_1 vrs P_3	97.5%	98.0%	99.4%	100%
P_2 vrs P_3	100%	99.4%	100%	100%
SNR-10				
P_1 vrs P_2	18.7%	48.0%	34.7%	64.0%
P_1 vrs P_3	16.0%	57.3%	50.7%	66.7%
P_2 vrs P_3	14.7%	45.3%	30.7%	58.7%
SNR-5				
P_1 vrs P_2	46.7%	78.7%	81.3%	77.3%
P_1 vrs P_3	44.0%	81.3%	78.7%	80.0%
P_2 vrs P_3	25.3%	74.7%	81.3%	93.3%
SNR 0				
P_1 vrs P_2	62.7%	82.7%	86.7%	92.0%
P_1 vrs P_3	70.7%	88.0%	89.3%	92.0%
P_2 vrs P_3	61.3%	96.0%	96.0%	97.3%
SNR 5				
P_1 vrs P_2	84.0%	92.0%	90.7%	96.0%
P_1 vrs P_3	85.3%	92.0%	94.7%	98.7%
P_2 vrs P_3	94.7%	97.3%	98.7%	100%
SNR 10				
P_1 vrs P_2	88.0%	88.0%	93.3%	97.3%
P_1 vrs P_3	90.7%	93.3%	98.7%	98.7%
P_2 vrs P_3	98.7%	98.7%	98.7%	100%

For example, in Table 1, it can be observed how Babble noise affects the performance of the classifiers. It can be established that noise levels higher than SNR0 (such as SNR-5 and SNR-10) produce accuracy levels above 50%. In other words, they affect several classifiers even below the level of random identification.

Among the classifiers used, SVM stands out as one that maintains the most stable accuracy percentage at all noise levels, although it is also severely affected at the highest noise levels.

Figure 2 verifies that the KNN classifier is the one that is affected the most by Babble noise, while SVM is the one presenting the highest level of robustness. In terms of classification of Clean footstep sounds, the accuracy percentage of all classifiers is very similar.

The results obtained by adding Office noise are shown in Table 2. For this kind of noise, it can be observed that, compared with the previous case, the performance of the classifiers is affected in a differing way. For example, the accuracy of KNN decreases very quickly, while the accuracy of SVM and Naive Bayes is affected more lightly. In the case of Random Forest, it is until SNR-10 that the accuracy drops considerably.

Table 2. Accuracy of classifiers for different levels of Office Noise.

Clean				
	KNN	N. Bayes	R. Forest	SVM
P_1 vrs P_2	96.6%	96.4%	98.2%	98.1%
P_1 vrs P_3	97.5%	98.0%	99.4%	100%
P_2 vrs P_3	100%	99.4%	100%	100%
SNR-10				
P_1 vrs P_2	7.2%	50.0%	41.6%	67.6%
P_1 vrs P_3	17.0%	51.6%	47.2%	72.8%
P_2 vrs P_3	4.4%	35.2%	22.8%	56.8%
SNR-5				
P_1 vrs P_2	25.2%	80.0%	84.4%	88.8%
P_1 vrs P_3	31.6%	77.2%	83.2%	88.0%
P_2 vrs P_3	16.4%	62.0%	72.0%	85.2%
SNR 0				
P_1 vrs P_2	58.0%	87.6%	90.8%	94.8%
P_1 vrs P_3	59.2%	88.0%	94.8%	98.4%
P_2 vrs P_3	51.6%	86.8%	95.6%	99.6%
SNR 5				
P_1 vrs P_2	86.8%	92.8%	96.8%	97.2%
P_1 vrs P_3	84.0%	95.2%	98.8%	99.2%
P_2 vrs P_3	90.8%	97.2%	99.2%	99.6%
SNR 10				
P_1 vrs P_2	96.8%	98.0%	97.6%	98.8%
P_1 vrs P_3	96.0%	95.6%	99.2%	99.6%
P_2 vrs P_3	99.6%	99.2%	100%	100%

Figure 2b shows how having higher noise levels, as in the case of SNR-5 and SNR-10, the performance of the KNN classifier decreases significantly, compared to classifiers such as Naive Bayes and SVM where the performance decreases less. It can also be observed that the performance of the classifiers was affected similarly to that obtained with Babble noise, where the accuracy of the KNN, Naive Bayes, and Random Forest classifiers is below 50% for noise levels higher than SNR0.

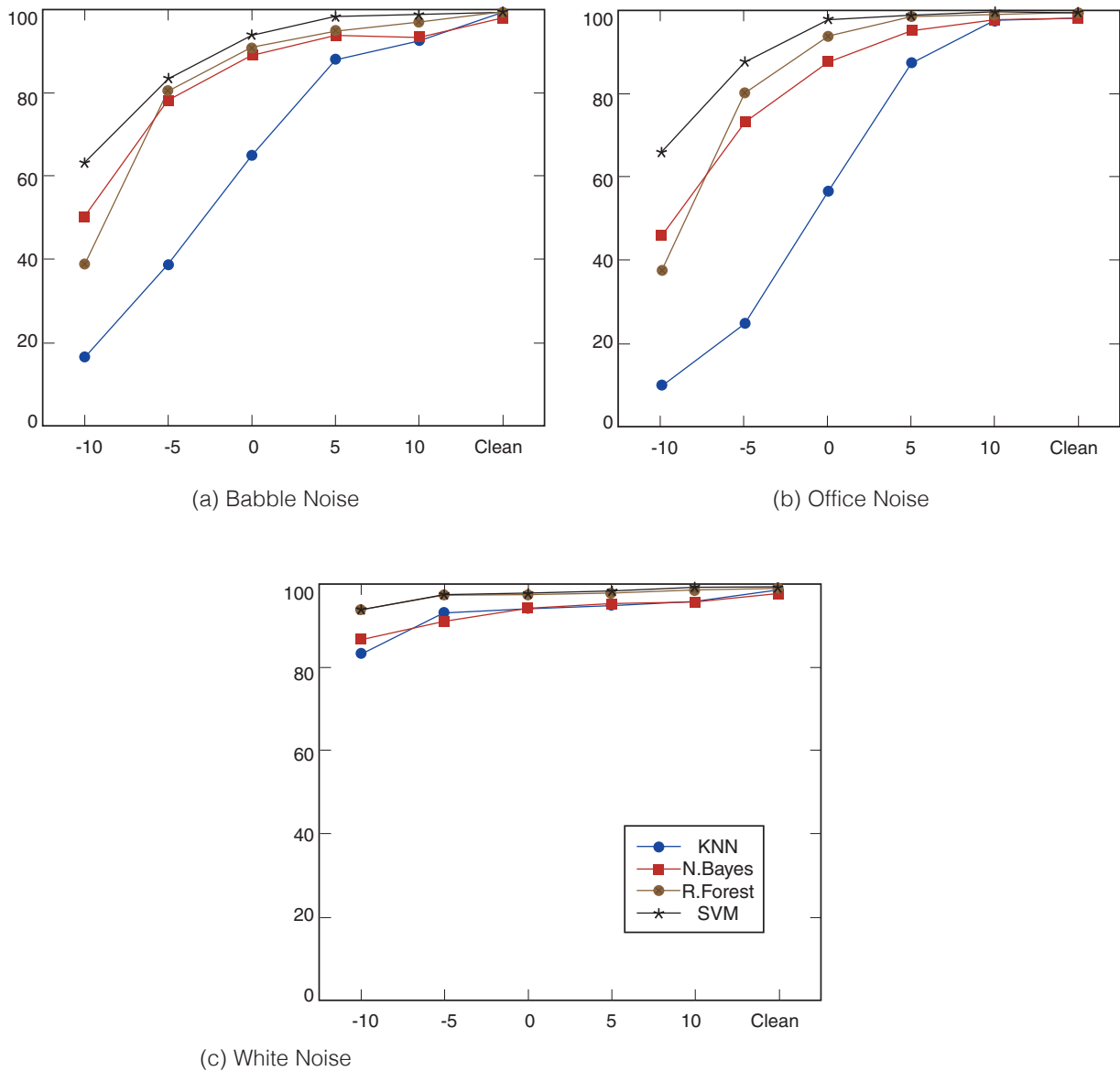


Figure 2. Accuracy of classifiers as a function of SNR (Mean of the three comparisons).

Finally, Table 3 shows the results for White noise. In these results, it can be seen that White noise affects classifiers to a lesser extent than the Babble and Office noises do. It can also be established that, in this case, White noise affects the KNN and Naive Bayes classifiers slightly more, this difference is more pronounced as the noise level increases.

None of the classifiers decrease their performance considerably; however, the ones that remain more constant are Random Forest and SVM compared to KNN and Naive Bayes.

Table 3. Accuracy of classifiers for different levels of White Noise.

Clean				
	KNN	N. Bayes	R. Forest	SVM
P_1 vrs P_2	96.6%	96.4%	98.2%	98.1%
P_1 vrs P_3	97.5%	98.0%	99.4%	100%
P_2 vrs P_3	100%	99.4%	100%	100%
SNR-10				
P_1 vrs P_2	93.6%	94.0%	98.0%	97.6%
P_1 vrs P_3	78.0%	81.6%	91.6%	92.8%
P_2 vrs P_3	79.2%	85.2%	92.4%	92.0%
SNR-5				
P_1 vrs P_2	96.8%	98.0%	99.6%	98.0%
P_1 vrs P_3	88.8%	82.0%	94.0%	96.4%
P_2 vrs P_3	94.0%	94.0%	98.8%	98.8%
SNR 0				
P_1 vrs P_2	99.2%	98.4%	98.8%	99.6%
P_1 vrs P_3	88.0%	87.6%	94.4%	95.2%
P_2 vrs P_3	95.6%	97.6%	97.6%	99.2%
SNR 5				
P_1 vrs P_2	99.2%	98.4%	99.2%	100%
P_1 vrs P_3	88.0%	88.8%	94.8%	95.6%
P_2 vrs P_3	98.0%	99.2%	100%	100%
SNR 10				
P_1 vrs P_2	98.0%	97.6%	99.2%	100%
P_1 vrs P_3	91.6%	90.4%	96.8%	98.4%
P_2 vrs P_3	98.8%	98.8%	100%	100%

Figure 2c confirms that White noise affects the KNN and Naive Bayes classifiers at high noise levels (SNR-10 and SNR-5) to a greater extent. In addition, it is observed that the performance of KNN and Naive Bayes is very similar, SVM and Random Forest also present a very similar behavior, the latter two being the ones with the best performance. As for the Clean steps, the classifiers have a similar accuracy.

Conclusions

In this work, a first study was presented on the use of the sound of steps captured with a single remote omni-directional microphone, as a means of biometric identification. Since an application of this type implies its use in a real environment, where there is always noise pollution, the main interest has been to analyze how much additive noise conditions can affect the identification of a person using the sound of their steps. For this, a database of step sounds recorded with controlled environmental conditions was developed. The binary classification in pairs of participating individuals was also used.

The results show a different affectation in the classifiers, for which those that can be considered simpler, such as KNN, are greatly affected as the noise level increases, compared to others such as SVM and Naive Bayes that have a greater robustness. The type of noise also affects

the affectation in a different way, where white noise presents little affectation in the identification of the person. In contrast, a natural noise such as the sounds of an office affects the classifiers much more.

As future work, the results of this work can be expanded in several directions, such as in the use of sound as a biometric identifier for a particular person with respect to a broader set of individuals and the impact of noise in this case. Noise reduction methods can also be tested for the developed database and to see if it allows for filtered sound to be considered as a means of biometric identification, in the case of favorable results. Finally, the possibility of improving the classification can be done using methods such as mixing experts and other classifiers based on deep learning.

References

- [1] Rodríguez, Rubén Vera, et al. "Footstep recognition for a smart home environment." *International Journal of Smart Home 2.2* (2008): 95-110.
- [2] Pedotti, Antonio. "Simple equipment used in clinical practice for evaluation of locomotion." *IEEE Transactions on Biomedical Engineering 5* (1977): 456-461.
- [3] Addlesee, Michael D., et al. "The ORL active floor [sensor system]." *IEEE Personal Communications 4.5* (1997): 35-41.
- [4] Rodríguez, Rubén Vera, Nicholas WD Evans, and John SD Mason. "Footstep recognition." *Encyclopedia of Biometrics*; Li, SZ, Jain, AK, Eds.; Springer: Boston, MA, USA (2015): 693-700.
- [5] Rodríguez, Rubén Vera, et al. "An experimental study on the feasibility of footsteps as a biometric." *2007 15th European Signal Processing Conference. IEEE, 2007.*
- [6] Vera-Rodríguez, Rubén, et al. "Analysis of time domain information for footstep recognition." *International Symposium on Visual Computing. Springer, Berlin, Heidelberg, 2010.*
- [7] Mason, James Eric, Issa Traoré, and Isaac Woungang. "Gait Biometric Recognition." *Machine Learning Techniques for Gait Biometric Recognition. Springer, Cham, 2016. 9-35.*
- [8] Connor, Patrick, and Arun Ross. "Biometric recognition by gait: A survey of modalities and features." *Computer Vision and Image Understanding 167* (2018): 1-27.
- [9] Vera-Rodríguez, Rubén, et al. "A large scale footstep database for biometric studies created using cross-biometrics for labelling." *2008 10th International Conference on Control, Automation, Robotics and Vision. IEEE, 2008.*
- [10] Shoji, Yasuhiro, Takashi Takasuka, and Hiroshi Yasukawa. "Personal identification using footstep detection." *Proceedings of 2004 International Symposium on Intelligent Signal Processing and Communication Systems, 2004. ISPACS 2004. IEEE, 2004.*
- [11] Hori, Yuki, Takahiro Ando, and Akira Fukuda. "Personal Identification Methods Using Footsteps of One Step." *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC). IEEE, 2020.*
- [12] Giannakopoulos, Theodoros. "pyaudioanalysis: An open-source python library for audio signal analysis." *PLoS one 10.12* (2015): e0144610.