





Dhinaharan Nagamalai  
Natarajan Meghanathan (Eds)

# Computer Science & Information Technology

6<sup>th</sup> International Conference of Advanced Computer Science & Information  
Technology (ACSIT 2018) May 26~27, 2018, Dubai, UAE



**AIRCC Publishing Corporation**

## **Volume Editors**

Dhinaharan Nagamalai,  
Wireilla Net Solutions, Australia  
E-mail: dhinthia@yahoo.com

Natarajan Meghanathan,  
Jackson State University, USA  
E-mail: nmeghanathan@jsums.edu

ISSN: 2231 - 5403

ISBN: 978-1-921987-86-1

DOI : 10.5121/csit.2018.80801 - 10.5121/csit.2018.80810

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

## Preface

The 6<sup>th</sup> International Conference of Advanced Computer Science & Information Technology (ACSIT 2018) was held in Dubai, UAE during May 26~27, 2018. The 6<sup>th</sup> International Conference on Information Technology in Education (ICITE 2018) and The 6<sup>th</sup> International Conference on Signal Image Processing and Multimedia (SIPM 2018) was collocated with The 6<sup>th</sup> International Conference of Advanced Computer Science & Information Technology (ACSIT 2018). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ACSIT-2018, ICITE-2018, SIPM-2018 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ACSIT-2018, ICITE-2018, SIPM-2018 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ACSIT-2018, ICITE-2018, SIPM-2018.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai  
Natarajan Meghanathan

## Organization

### General Chair

David C. Wyld  
Jan Zizka

Southeastern Louisiana University, USA  
Mendel University in Brno, Czech Republic

### Program Committee Members

Abdulrahman A.A.Emhemed  
Adam Taylor  
Ahmed Abdou  
Akram Abdelqader  
Alaa Hamami  
Ali El-Zaart  
Almir Pereira Guimaraes  
Amir Salarpour  
Amiya Kumar TRIPATHY  
Anamika Yadav  
Atallah Mahmoud AL-Shatnawi  
Ayman EL-SAYED  
Azeddine Chikh  
Belete Biazen  
Benaissa Mohamed  
Carlo Sau  
Chaker LARABI  
Chin-chen Chang  
Christian Mancas  
Chuanzong Zhang  
Dac-Nhuong Le  
Debjani Chakraborty  
Debotosh Bhattacharjee  
El Miloud Ar Reyouchi  
Elena Battini Sonmez  
Emad Awada  
Essam Halim Houssein  
Fatih Korkmaz  
Gabor Kiss  
Gabriel Badescu  
Gebeyehu  
Goran Bidjovski  
Gullanar M Hadi  
Guo-Shiang Lin  
Hacer Yalim Keles  
Hamid Alasadi  
Hayet Mouss

College of Technical Sciences. Bani Walid, Libya  
Trinity College Dublin, Ireland  
Al-Quds University, Palestine  
AL-Zaytoonah University of Jordan, Jordan  
Princess Sumaya University for Technology, Jordan  
Beirut Arab University, Lebanon  
Federal University of Alagoas, Brazil  
Bu-Ali Sina University, Iran  
Edith Cowan University, Australia  
National Institute of Technology, India  
Al al-Byte University, Jordan  
Menoufia University, Egypt  
University of Tlemcen, Algeria  
Bahir Dar Institute of Technology, Ethiopia  
Univ Ctr Of Ain Temouchent, Algeria  
Universita degli Studi di Cagliari, Italy  
Universite de Poitiers , France  
Feng Chia University, Taiwan  
Ovidius Univesrity, European Union  
Aalborg University, Denmark  
Haiphong University, Vietnam  
Indian Institute Of Technology, India  
Jadavpur University, India.  
Abdelmalekessaadi University, Morocco  
Istanbul Bilgi University, Istanbul.  
Applied Science University, Jordan  
Minia University Egypt  
Cankiri Karatekin University, Turkey  
Obuda University, Hungary  
University of Craiova, Romania.  
Bahir Dar University, Ethiopia.  
International Balkan University, Macedonia  
Salahaddin University, Iraq  
Da-Yeh University, Taiwan  
Ankara University, Turkey  
Basra University, Iraq  
Batna Univeristy, Algeria

Hector Migallon	Miguel Hernandez University, Spain
Hongzhi	Harbin Institute of Technology, China
Ibtihel Nouira	Technologie and Medical Imaging Laboratory, Tunisia
Ireneusz Kubiak	Military Communication Institute, Poland
Isa Maleki	Islamic Azad University, Iran.
Ivan Izonin	Lviv Polytechnic National University, Ukraine.
Ivo Pierozzi Junior	Embrapa Agricultural Informatics, Brazil
Jafar Mansouri	Ferdowsi University of Mashhad, Iran
Jamal El Abbadi	Mohammadia V University Rabat, Morocco
Jun Zhang	South China University of Technology, China
Keneilwe Zuva	University of Botswana, Botswana
Klimis Ntalianis	Athens University of Applied Sciences, Greece
Kurd gift	IMAM University, Saudi Arabia
Madya Dr. Mohammad Bin Ismail	Universiti Malaysia Kelantan, Malaysia
Marco Furini	Universita Di Modena E Reggio Emilia, Italy
Mario Henrique Souza Pardo	University of Sao Paulo, Brazil
Marius CIOCA	Lucian Blaga University of Sibiu, Romania
Maryam hajakbari	Islamic Azad University, Iran
Miguel A. Rodriguez-Hernandez	ITACA Universitat Politecnica de Valencia, Spain
Mimoun Hamdi	Ecole Nationale d'Ingenieurs de Tunis, Tunisia
Mohamed B. El_Mashade	Al_Azhar University, Egypt
Mohamed HAMLICH	Hassan II University, Morocco
Mohamed Khayet	University Complutense of Madrid, Spain
Mohamed SENOUCI	Universite d'Oran 1 Ahmed Ben Bella, Algeria
Mohamed-Khireddine Kholadi	Echahid Hamma Lakhdar University, Algeria
Mohamedmaher Benismail	King saud University, Saudi Arabia
Mohammad Rawashdeh	University of Central Missouri, United States
Mostafa Ashry	Alexandria University, Egypt
Mourchid mohammed Ibn	Tofail University Kenitra, Morocco
Murat Tolga OZKAN	Gazi University Fakulty of Technology, Turkey
Naveed Ahmed	University of Sharjah, UAE
Noha Abdul Kareem Alnazzawi	Jubail University College, Saudi Arabia.
Noura Taleb	Badji Mokhtar University, Algeria
Ouafa Mah	Ouargla University, Algeria
Paulo Roberto Martins de Andrade	University of Regina, Canada
Petr Hajek	University of Pardubice, Czech Republic
Petrellis N	TEI of Thessaly, Greece
Prakash Duraisamy	University of Central Missouri, United States
Roberto De Virgilio	Roma Tre University, Italy
Ruchi Tuli	Jubail University College, Saudi Arabia.
Saban Gulcu	Necmettin Erbakan University, Turkey
Salem Nasri	Qassim University, Saudi Arabia
Shoeib Faraj	Institute of Higher Education of Miaad, Iran
Taeghyun Kang	University of Central Missouri, United States
TCHIOTSOP Daniel	University of Dschang, Cameroon
Tesfahunegn M/Mengistu	Bahir Dar Institute of Technology, Ethiopia
Tiendrebeogo Telesphore	Universite of Nazi Boni, Burkina Faso
Wee kuok kwee	Multimedia University , Malaysia
Xuechao Li	Auburn University, USA

## **Technically Sponsored by**

**Computer Science & Information Technology Community (CSITC)**



**Networks & Communications Community (NCC)**



**Soft Computing Community (SCC)**



## **Organized By**



**Academy & Industry Research Collaboration Center (AIRCC)**



## TABLE OF CONTENTS

### **6<sup>th</sup> International Conference of Advanced Computer Science & Information Technology (ACSIT 2018)**

**Global Music Asset Assurance Digital Currency : A DRM Solution for Streaming Content Using Block Chain** ..... 01 - 11  
*Ahmed Gomaa*

**Detection of Algorithmically Generated Malicious Domain** ..... 13 - 32  
*Enoch Agyepong, William J. Buchanan and Kevin Jones*

**Two Discrete Binary Versions of African Buffalo Optimization Metaheuristic** ..... 33 - 46  
*Amira GHERBOUDJ*

**Intelligent Electronic Assessment for Subjective Exams** ..... 47 - 63  
*Alla Defallah Alrehily, Muazzam Ahmed Siddiqui and Seyed M Buhari*

**Dynamic Phone Warping - A Method to Measure the Distance Between Pronunciations** ..... 65 - 73  
*Akella Amarendra Babu and Ramadevi Yellasiri*

**A Survey on Question Answering Systems : The Advances of Fuzzy Logic**..... 75 - 93  
*Eman Mohamed Nabil Alkholy, Mohamed Hassan Haggag and Constantine's Koutsojannis*

### **6<sup>th</sup> International Conference on Information Technology in Education (ICITE 2018)**

**Promoting Student Engagement Using Social Media Technologies**..... 95 - 105  
*Mohammad Alshayeb*

**Moving from Waterfall to Agile Process in Software Engineering Capstone Projects**..... 107 - 114  
*Mohammad Alshayeb, Sajjad Mahmood and Khalid Aljasser*

### **6<sup>th</sup> International Conference on Signal Image Processing and Multimedia (SIPM 2018)**

**4D Automatic Lip-Reading for Speaker's Face Identification**..... 115 - 126  
*Adil AbdUlhur AboShana*

<b>Analysis of Land Surface Deformation Gradient by Dinsar.....</b>	<b>127 - 136</b>
<i>Karima Hadj-Rabah, Faiza Hocine, SawsenBelhadj-Aissa and Aichouche Belhadj-Aissa</i>	

# GLOBAL MUSIC ASSET ASSURANCE DIGITAL CURRENCY: A DRM SOLUTION FOR STREAMING CONTENT USING BLOCKCHAIN

Ahmed Gomaa

Department of Operations and Information Management,  
University of Scranton, Scranton, USA

## ABSTRACT

*The amount of piracy in the streaming digital content in general and the music industry in specific is posing a real challenge to digital content owners. This paper presents a DRM solution to monetizing, tracking and controlling online streaming content cross platforms for IP enabled devices. The paper benefits from the current advances in Blockchain and cryptocurrencies. Specifically, the paper presents a **Global Music Asset Assurance (GoMAA)** digital currency and presents the **iMediaStreams Blockchain** to enable the secure dissemination and tracking of the streamed content. The proposed solution provides the data owner the ability to control the flow of information even after it has been released by creating a secure, self-installed, cross platform reader located on the digital content file header. The proposed system provides the content owners' options to manage their digital information (audio, video, speech, etc.), including the tracking of the most consumed segments, once it is release. The system benefits from token distribution between the content owner (Music Bands), the content distributor (Online Radio Stations) and the content consumer(Fans) on the system blockchain.*

## KEYWORDS

*Blockchain, Cryptocurrency, Digital Rights Management, Public Key, Private Key*

## 1. INTRODUCTION

### 1.1 Streaming Content Problem

The online streaming music in the United States has been increasing in the past several years. It currently accounts for 65% of the music online share [1]. The multibillion dollar industry is constantly faced with intellectual rights infringement. For instance, in January, 2018, Spotify, a music streaming company was sued by Wixen Music Publishing Inc. for allegedly using thousands of songs, without a license and compensation to the music publisher [2].

There are two main music royalty collecting societies in the USA: the American Society of Composers, Authors and Publishers (ASCAP) and, Broadcast Music Inc. (BMI), with hundreds of thousands of members each. If two artists collaborated on the same music album, but are subscribed to different royalty collecting societies, they will receive different royalties, a fact that shows the discrepancy in how played streams are counted in different organizations. As of 2016,

ASCAP and BMI alone collect and disburse payments in the range of \$1.8 billion annually on behalf of hundreds of thousands of musicians for royalties around the world. It is not the actual value of the market. According to Institute for Policy Innovation (IPI) 2007 report, that it is costing the US economy more than 12 Billion dollars due to sound recording piracy in the US. In 2017, ASCAP and BMI announced the creation of a new comprehensive musical works database to increase ownership transparency in performing rights licensing that is expected to roll out at the end of 2018 [3]

The problem becomes more challenging when considering the online radios, and the Disk Jockeys (DJs) who are mixing music live and stream it online with audience listening around the world. The sale and distribution of media content using a digital medium provides simple and flexible production, consumption, and transmission of such content. However, it also reduces the efforts needed for unauthorized usage of this data. Thus, digital media content is more easily copied, distributed, or used in a manner not allowed by law or license agreement.

## 1.2 Blockchain Technology Overview

Blockchain enables secure peer to peer transactions [4]. There is a number of existing public blockchain platforms including Ethereum and Bitcoin Blockchain. Each platform has its own purpose. For instance, Bitcoin Blockchain enables a peer to peer cash system to allow online payments and allows for tracking ownership of the digital currency. On the other hand, Ethereum platform focuses on running the programming code of any decentralized application [5]. To incentivise and assure the decentralization concept, each blockchain platform rewards the platform participants, called miners, with coins, either Bitcoins on the Bitcoin Blockchain or Ether on the Ethereum blockchain. It may be viewed as a compensation for the work done. Each block on the blockchain includes a number of transactions. The blockchain platforms are designed for the entire network to consume electricity proportionate to the amount of coins given to the miners. Once a block is added, all miners compete to solve the computational problem for the new block. On the bitcoin blockchain, a new block is created every 10 minutes on average, while on the Ethereum blockchain, a new block is created every 15 seconds on average. At any given day, the amount of money spent by the entire network on electricity will be proportional to the amount of money gained by those who find the correct answer first. For instance, on a certain day in April 2018, on the Bitcoin Blockchain, 153 blocks were mined, with an average time of 8.72 minutes between every block, the 153 blocks contained 226,626 transactions. The estimated profits from the miners on that day was \$17,344,399.98, on the other hand 43% of that amount is estimated to be spent on the electricity by the entire network. In the Ethereum blockchain, instead of mining for bitcoin, miners work to earn Ether. Since Ethereum blockchain is designed to run programming code on decentralized application, its protocol can be used as a tool for self-operating computer programs that automatically executes when specific conditions are met.

## 1.3 Public Key / Private Key infrastructure

Both Bitcoin and Ethereum blockchain implementations use public key / Private Key infrastructure, where it is used in digital wallet creation and transaction generation and verification.

For instance, to generate a Bitcoin wallet, the digital wallet software will generate a new private key and a corresponding public key. These keys are later used by the owner of the wallet to send and receive coins on the platform. The concept is that someone may digitally sign an item, to confirm they are allowing an action to take place with a private key. In the Bitcoin implementation, it is sending money. Once a transaction is signed by the owner and sent to miners to be added to the blockchain, after the proof of work is done by the miners and a new block is added to the blockchain, the money is sent from one wallet to another.

The paper attempts to benefit from the proof of concept manifested in public key / private key infrastructure implemented on the Ethereum blockchain along with the capability of running decentralized programming applications on the blockchain to address the piracy problem on streamed online content on a new blockchain network, named “iMediaStreams Blockchain”. iMediaStreams Blockchain is a decentralized solution to music publishing houses as it addresses the process of validating the authenticity of the music streams.

#### **1.4 Current Solutions and their problems:**

Lawmakers recognized the growing need to protect digital media and enacted the US Digital Millennium Copyright Act (DMCA) [6, 7] to protect property rights. One approach to curbing the proliferation of illegal activity surrounding digital media content is to incorporate a form of Digital Rights Management (DRM) into the digital content. DRM can be used to detect and verify ownership of data and to control access to the data in accordance with a policy determined by the content creator or distributor. A further approach frequently incorporated in a DRM system is to embed a digital watermark in the digital media file. A digital watermark is information that is generated and interspersed in the data of the digital media file but cannot be perceived by the audience of the digital media file [8, 9, 10]. For example, to a listener, a digital audio file that contains a watermark would sound identical to the digital audio file without the watermark. However, an examination of the data (e.g., by media player software) can detect (i.e., extract) the watermark to determine if the file has been modified in some way. Possible modifications that could alter the watermark include compression of the data [11-19], cropping an image or video or attempted removal of the watermark. Watermarking can also be used to fingerprint a file; such that different recipients receive differently watermarked content. Thus, by examining the watermark, the proper owner of a file can be determined and any tampering with the file can be detected. Attempts to address the DRM in streaming music are presented in [20], [21], and [22] where the main difference in this paper lies in the usage of the blockchain technology and having a digital wallet within the browser plugin. [23] Introduced the concept of using Blockchain in controlling digital content, but did not allow for cross platform rendering and access policy control on the media file. [24] Presents an access control policy on the blockchain. The main difference in the paper in contrast with the recent publications and patents is in the way this paper includes the access control on the file header instead and the build in player, an architectural difference that allows for the destruction of the file even after being sent out. Current commercial DRM solutions focus on watermarking digital media content to indicate ownership of a specific copy of the content and ways to track and prevent unauthorized reproductions and distributions. One such solution is Windows Media Rights Manager (i.e., windows media DRM 10). However, this solution is limited to a Windows environment and is not compatible or accessible on other computing platforms. Furthermore, it provides only a single layer of security, which if defeated can expose all content distributed with the system.

Currently, each major legal download services uses its own proprietary DRM algorithm, limiting which portable playback devices consumers can use with any given system. For example, music bought on iTunes can only be played on devices with an iTunes software installed. Still, the music industry is in need to protect its constituent's rights, while promoting the music industry by not being constrained by the hardware or software played. Further, it is important to track and control the released files that are aired online. Being able to provide a control mechanism on the released songs, and a transparent reporting capability is highly valued by reporting agencies as well as by musicians and production houses.

Due to the use of such digitized media, new tools are needed by the owners of Intellectual Property and digital content to assert their rights and to prevent unauthorized usage. Content owners are interested in a number of aspects including

- Means to indicate ownership on online contents for monetization.
- Tracking digital content consumptions.
- Controlling the dissemination of the content already released.

What is needed in the art is a platform-independent DRM system for monetizing, tracking, and controlling the usage of digital media content and providing robust multi-level security to prevent the subversion of the protection system.

## 2. PROBLEM STATEMENT

Given the current state of piracy in the online music industry, and the current advancement in blockchain platforms, the paper is addressing how to monetize, track and control after dissemination online streamed content. This paper proposes a Digital Rights Management solution for the stated problems using blockchain and cryptocurrency technology

## 3. PROPOSED SOLUTION

### 3.1 System Architecture

To address the tracking problem of music, the relationship between the music bands, online streaming portals and the fans needs to be addressed. The proposed solution depicted in figure 1 promotes the reward of the fans for listening to their favourite music bands if the fans opted to do so. This is needed to incentivise the fans to participate in the system. The first step for the music bands is to generate their content into a portable secure format, that is self-rendering and cross platform. For streaming portals to include that content, they will need to use a Global Music Asset Assertion (GoMAA) token to allow them to stream the content on their portals. As the Fans tune in the online streaming portals, the fans will be presented with the option to install a browser plugin which includes a digital wallet, the fans will collect GoMAA tokens as they consume the content. The fans start collecting rewards for every streaming content they consume; the online radio stations provide those rewards in GoMAA tokens. All token transactions are recorded on the iMediaStreams blockchain, which includes the streaming site, the content played, and the listeners' wallet hashed address. For online radio stations, number of songs streamed will be recorded on the blockchain. For Fans, if they opted to use their digital wallet to collect coins, the number of songs they listened to will be recorded on the blockchain, using the fans wallet hashed address. Having all the information on the blockchain addresses the royalty discrepancy problem as the information is accessed by all while preserving the privacy of the streaming stations, listeners and music bands. The following sections explain the system components in details.

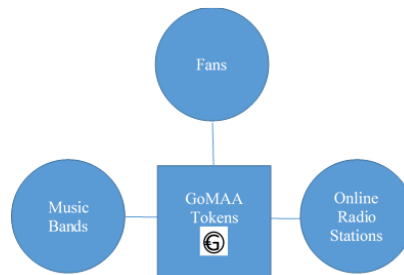


Figure 1: DRM Global Music Asset Assertion Token (GoMAA)

### 3.1.1 The Global Music Assets Assurance (GoMAA) Token on iMediaStreams Blockchain

The iMediaStreams Blockchain ecosystem is a safe environment for the buying, loaning, renting and selling of media streams. This environment is securely accessed through the GoMAA Tokens as the only means of exchange within the platform.

### 3.1.2 The Coin

The GoMAA Coin token is a cryptographic token, created to be exchanged for digital streaming content. It is the only mean of exchange on the iMediaStreams Blockchain to give all participants access to value they seek, whether they are the music producers, the online radio stations or the fans. The token is of fixed supply, fractionally divisible and non-inflationary over the long term. GoMAA Coins may be exchanged on the major cryptocurrency exchanges.

### 3.1.3 Protocol

The GoMAA Coin token is using an ERC20 [25] token protocol on the iMediaStreams blockchain. This token interface will enable the issuance of the GoMAA Coin token to be integrated and utilized within iMediaStreams ecosystem, which includes a Points Reward System for the media consumers. The tokens will be used by the online radio stations to buy streaming rights from Music Bands or to encourage loyal fans to explore more music or a combination of both. Every Music Band and Fan will set their value based on information provided about their value as determined by the iMediaStreams blockchain public record. A Fan may share preferences, their network of friends and followers, or even offer to share unused computing power or storage, just to name a few. All these make the Fan more attractive to both streaming radio stations and Music Bands, which will make them being valued at a higher token share. Music Brands may provide more information about their fan base, the venues, and their tours, for commanding a higher valuation for their content in terms of token required to stream their content.

iMediaStreams Blockchain creates an ecosystem where all participants are compensated through the usage of GoMAA tokens based on smart contract fulfilment, which creates a DRM environment where everyone benefits.

To be able to use the tokens on the blockchain network, Figure 2 shows the overall architecture of the DRM solution

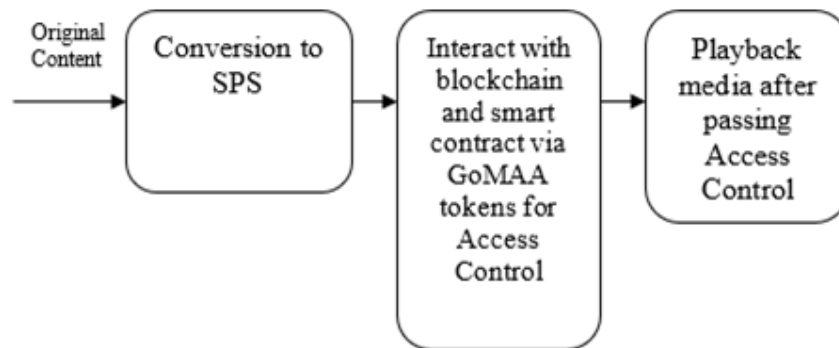


Figure 2: The overall architecture of the DRM solution

The details of the Conversion to SPS format are shown in Figure 3 and the format of the SPS is shown in Figure 4. Figure 5 shows the playback process at the receiver end whereby the content released to another party undergoes the Blockchain validation before the final rendering.

### 3.2 Secure Portable Streaming Format (SPS)

The first step is in converting the media file in a high level proprietary format that is both secure and portable. In addition, the content needs the options for tracking and dissemination control once released. Figure 3 shows this conversion process which converts the original content into a Secure Portable Streaming format (SPS).

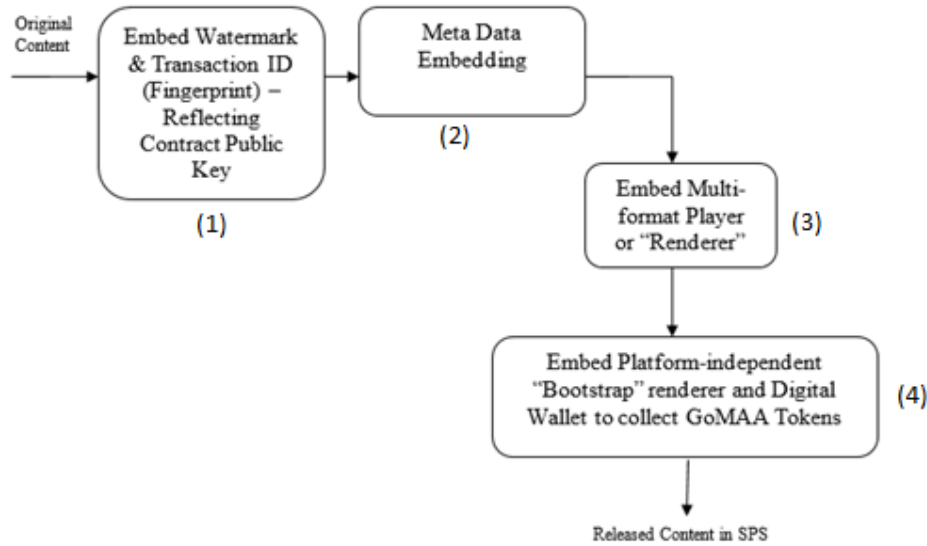


Figure 3: Conversion of original content to SPS

1. The initial content is encoded to insert a Watermark and Transaction ID (fingerprint). This is utilized when validating the transaction on the blockchain to verify privileges and rights. The fingerprint is the smart contract public key of the streaming content. Every streaming content will have its own corresponding smart contract. Prior to creating the contract address, the original content file itself is augmented with a high frequency hertz wave ( $> 20,000$  Hz) to be outside of humans' ears capabilities, but act as a signature to the audio file, without impacting the quality of the audio. The fingerprint can be added to the audio file in intervals determined by the content owner.
2. In addition to having access control on the smart contract of the audio file, the Meta Data structure is defined based on additional usage, protection and distribution constraints as specified by the content owners. For this, the paper uses a meta-data security mechanism based on the digital content owners' requirements for protection and tracking of their content. This security mechanism is integrated into the DRM system and characterizes as a second layer in the multi-level security information flow involving permission to stream and copy, with build in instruction for access rights, including permission to stream, copy, or transfer information, "time-to-live", number of viewings allowed, even self-destruction. The security policy within the metadata on the SPS file header articulates the number of tokens needed to consume the streamed content intervals. The consumed content is published on the blockchain.



3. The renderer for the online radio streaming server is only activated based on the available tokens on the radio station wallets. Those tokens will be used to pay the content owners after the streaming is done. The renderer for the fans will be activate according to the content smart contract. If tokens are required from the fans to listen to the content, it will not be rendered to them unless they have enough tokens to pay after receiving the required stream. If no tokens are required, the plugin will enable the rendering of the files.
  4. The amount and types of available tokens enables different activities. Token act as private keys that are needed. The public key represented as the streaming content id is known to everybody and the private or secret key is only known to the intended recipient who own the tokens.
- Once a content is streamed, the smart contract clears the transactions, and the GoMAA token passes from the streaming portal to the content owners, the fans, and the iMediaStreams blockchain. If the streaming of the content cannot be validated, it is presumed to be invalid, and the GoMAA token does not transfer. The final outcome is recorded and provable. This will directly lead to the conclusion of the reconciliation debate between different tracking organizations.

### 3.2 Data Format of the SPS

Figure 4 shows a sample bit-stream for the SPS format. The encryption key is embedded in the machine code (which is invisible) for different players such as Win 32/64, UNIX, Mac etc. The transaction ID (or equivalently fingerprint) which contains the privilege or rights information is part of the Encrypted code.

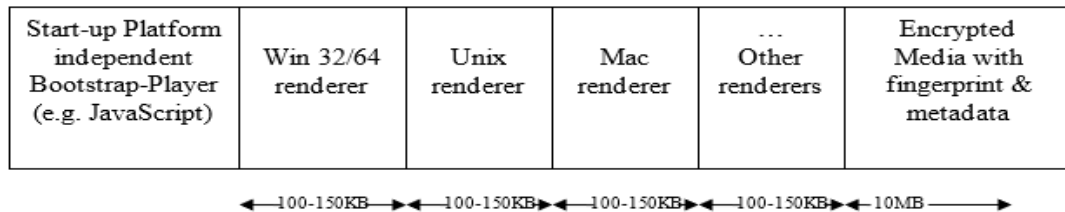


Figure 4: Sample format for SPS

The file can be self-rendered, given that a browser has a digital wallet plugin, which includes GoMAA tokens if required by the content owner and articulated in the media smart contract.

### 3.3 Playback/Rendering of released content or media

The renderer or player is a cross platform plugin installed on the fans' browsers. The plugin act as a digital wallet and as a media render, as depicted in figure 5 below.

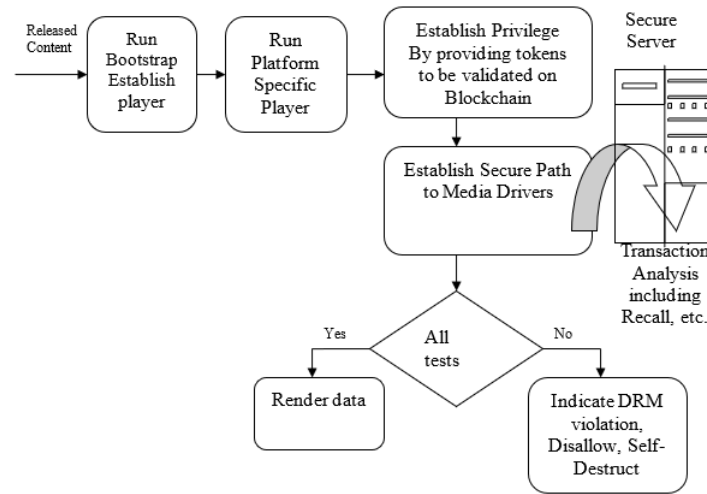


Figure 5: Playback or Rendering Process

The renderer or player aims for a self-extracting auto-executing player and digital wallet which is compatible across different browsers. This plugin prepares the content for the next step in the process, namely, the Access Control policy defined on the file header/. In order for the player to be able to control the original content after dissemination, invocation codes associated with that data are incorporated and installed on the player, where the required parameters include the predefined limitations on the data. Each time an action is taken on the content via the player (Play, FF, RW, etc.) the associated invocation code is activated and the action is counted and compared versus the original policy set at the server side. If the action is satisfied, the appropriate code is sent to the player to enable the file to play. The player consists of a *bundle* of optimized and compact platform specific players (each being machine specific binary code) along with a portable bootstrap script (for which a Java script model is used).

## 4. USE CASES

### 4.1 Buying Streaming content rights Using iMediaStreams Blockchain

iMediaStreams Blockchain provides an open ledger solution to address the discrepancy in monetising, tracking and controlling online streams. It provides music professionals with an environment where they release their music with 100% confidence that their high quality original content will be traced all the time. The transaction, once validated, is recorded to the blockchain and there is no longer any confusion or debate as to the transaction and reconciliation.

The proposed iMediaStreams Blockchain is built on an ERC-20 standard, iMediaStreams Blockchain utilizes smart contracts to enforce this transaction. iMediaStreams Blockchain, in its simplest form, is effectively a white list of legitimate online music disseminators and retailers. The smart contract quickly assesses if the song played on a domain validated by its record on the blockchain. If it doesn't appear on the blockchain and is a new domain to the environment, iMediaStreams Blockchain acts as a validation point using its data relationships to assess the played stream. Once the played stream is validated, the smart contract clears the transaction and the GoMAA Token passes from the content disseminator to the content owner. If the played stream cannot be validated, it is presumed the played stream is invalid and the GoMAA Token does not move. The final outcome is recorded and provable and any potential debate over reconciliation disappears between content owners and content disseminators. Since transactions

are validated in a short period of time, iMediaStreams Blockchain creates another change for music industry. Payment cycles can be shortened to the time it takes the smart contract to complete. If both parties desire, payment is nearly instantaneous. The option of reducing payment cycles from a few months to minutes is possible. The tracking agency collects a royalty, payable in GoMAA Tokens, on these transactions. These commissions are charged as a percentage of the transaction value.

#### **4.2 iMediaStreams Blockchain to Compensate Listeners:**

Listeners of songs are drawn to pirated content to save money. They can actually be compensated by the system if they are enjoying a higher quality song from their preferred bands. iMediaStreams Blockchain gives the content consumer ownership on how to use their tracked information, if any. The iMediaStreams Blockchain lets the listeners receive compensation. Upon consuming a digital asset, listeners receive GoMAA Tokens as a reward. These tokens can be used for many purposes within the iMediaStreams ecosystem ranging from free ad blocking to promotional offers from similar bands. Only by completing the value circle do all parties of the digital asset consumption transaction benefit. This is the value exchange missing from the online music industry today.

### **5. CONCLUSION**

The paper presents a solution for a number of problems in the DRM space. First, it presents an overarching proposition to music monetization, tracking, and controlling. The music monetization takes place by allowing the content owners to have their content accessed using the public/private key infrastructure within the iMediaStreams blockchain. Every time the content is streamed or downloaded, GoMAA tokens are exchanged according to the rules depicted on the content smart contract.

The Tracking of the music takes place by two means, the first is by querying the iMediaStreams blockchain, where information about how many time, each content is consumed is recorded. The second mean is by checking the security server that validates the access policy on the file header. The controlling part is placed as an invocation code of the player, requiring a confirmation from the Security server to render the content or even destroy it based on the security policy as depicted on the file header. In specific, the paper presents a system for generating and controlling access to copy-protected media files. The system includes a server having a processor and a computer readable medium encoding a server software program. The server software program is configured to encode and, encrypt the media content, store the resulting data in the digital media file, embed a transaction ID, or the smart contract public address, in addition to a user-access policy stored in the digital media file header, and embed a multi-format renderer in the digital media file. The multi-format renderer is configured to render the encrypted electronic file, and is further configured to generate an invocation code in response to a requested operation on the electronic file, retrieve the transaction ID associated with the electronic file, compare the invocation code to the user-access policy, and selectively respond to the requested operation based on a result of the comparison of the invocation code to the user-access policy. Future work includes tracking subsections used in streamed online content and how to create a valuation model for the content disseminated and the Fans using their information and influential relationships within the ecosystem.

In summary, the paper presents three main advantages:

1. A DRM system focusing on the secure delivery of entertainment and information to IP devices and television set-top boxes

2. A system that accounts for files consumption in digital streaming formats and for determining usage fees such as copyright royalties.
3. Allowing content distributors the freedom to pursue new revenue opportunities made possible by digital distribution by development and licensing copy protection, electronic licensing and rights management technologies

## REFERENCES

- [1] Datta, Hannes, George Knox, and Bart J. Bronnenberg. (2017) "Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery." *Marketing Science*
- [2] Carlos Micames. (2018) "SPOTIFY HIT WITH \$1.6 BILLION COPYRIGHT INFRINGEMENT LAWSUIT." American University Intellectual property brief,
- [3] Lauren Iossa, Cathy Nevins, Liz Fischer. (2017)"ASCAP & BMI Announce Creation Of A New Comprehensive Musical Works Database To Increase Ownership Transparency In Performing Rights Licensing" <https://www.ascap.com/press/2017/07-26-ascap-bmi-database>
- [4] Chahbaz, Ahmed. (2018). An Introduction to Blockchain and its Applications. With a Focus on Energy Management. diplom. de. ISBN 9783960677178
- [5] Wood, Gavin. (2014) "Ethereum: A secure decentralised generalised transaction ledger." *Ethereum Project Yellow Paper* 151 -1-32.
- [6] B.H. Turnbull, (2001) "Important Legal Developments Regarding Protection of Copyrighted Content Against Unauthorized Copying," *IEEE Communications Magazine*, pp. 92-100,
- [7] The Digital Millennium Copyright Act of 1998, US Copyright Office Summary, 1998, available at <http://www.copyright.gov/legislation/dmca.pdf>.
- [8] N. Sinha, (2000) "A Novel Watermarking Technique for digital images based on Adaptive Segmentation and Space-Frequency representation," *Proc. IEEE 2000 International Symposium on Information Theory and its Applications, ISITA 2000*, Vol. II, 972-975.
- [9] N. Sinha, (2002) "A New Digital Rights Management Platform for Digital Images," in the *Proceedings of IASTED International Conference on Communications, Internet and Information Technology (CIIT 2002)*, St. Thomas, USA, ISBN 0-88986-327-X, pp. 444-448
- [10] N. Sinha, (2005) "Secure Embedded Data Schemes for User Adaptive Multimedia Presentation", *Journal of Digital Information*, Volume 6, Issue 4, Article No. 350,
- [11] K. Brandenburg, G. Stoll, et al. (1992)"The ISO- MPEG-Audio Codec: A Generic-Standard for Coding of High Quality Digital Audio," in *92nd AES Convention*, 1992, Preprint no. 3336.
- [12] Marina Bosi et al., (1996) "ISO/IEC MPEG-2 Advanced Audio Coding," *101st Convention of the Audio Engineering Society*, Preprint n. 4382.
- [13] Mark Davis, (1993) "The AC-3 Multichannel Coder," *95th Convention of the Audio Engineering Society*, Preprint n. 3774.
- [14] J. D. Johnston, D. Sinha, S. Dorward, and S. R Quackenbush,(1996) "AT&T Perceptual Audio Coding (PAC)," in *AES Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. pp. 73-82.
- [15] *IEEE Transactions on Circuits and Systems for Video Technology* (2003):Special Issue on the H.264/AVC Video Coding Standard

- [16] D. Sinha and A. Ferreira (2005) "A New Broadcast Quality Low Bit Rate Audio Coding Scheme Utilizing Novel Bandwidth Extension Tools," 119th Convention of the Audio Engineering Society. Paper 6588.
- [17] A. Ferreira and D. Sinha, (2005) "A New Low-Delay Codec for Two-way High-Quality Audio-Communication," 119th Convention of the Audio Engineering Society, Paper 6572.
- [18] D. Sinha, A. Ferreira, and, D. Sen (2005) "A Fractal Self-Similarity Model for the Spectral Representation of Audio Signals," 118th Convention of the Audio Engineering Society, Paper 6467.
- [19] A. Ferreira and D. Sinha, (2005) "Accurate Spectral Replacement," 118th Convention of the Audio Engineering Society, Paper 6383.
- [20] Ahmed Gomaa, D. Sinha (2009), Cross-platform digital rights management providing multi-level security information flow tracking, US Patent App. 12/209,893, 200
- [21] Sahita, R., & Covington, M. J. (2014). U.S. Patent No. 8,689,349. Washington, DC: U.S. Patent and Trademark Office.
- [22] Soppera, A., & Burbridge, T. (2013). U.S. Patent No. 8,533,782. Washington, DC: U.S. Patent and Trademark Office.
- [23] Kishigami, J., Fujimura, S., Watanabe, H., Nakadaira, A., & Akutsu, A. (2015). The blockchain-based digital content distribution system. In Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on (pp. 187-190). IEEE.
- [24] Maesa, D. D. F., Mori, P., & Ricci, L. (2017). Blockchain based access control. In IFIP International Conference on Distributed Applications and Interoperable Systems (pp. 206-220). Springer, Cham.
- [25] Fabian Vogelsteller. 2015. ERC20. (2015). <https://github.com/ethereum/eips/issues/20>.

*INTENTIONAL BLANK*

# DETECTION OF ALGORITHMICALLY GENERATED MALICIOUS DOMAIN

Enoch Agyepong<sup>1</sup>, William J. Buchanan<sup>2</sup> and Kevin Jones<sup>3</sup>

<sup>1</sup>Cyber Operations Team, Airbus, Corsham, UK

<sup>2</sup>School of Computing, Edinburgh Napier University, Edinburgh, UK

<sup>3</sup>Cyber Operations Team, Airbus Group Innovations, Newport, Wales, UK

## ABSTRACT

*In recent years, many malware writers have relied on Dynamic Domain Name Services (DDNS) to maintain their Command and Control (C&C) network infrastructure to ensure a persistence presence on a compromised host. Amongst the various DDNS techniques, Domain Generation Algorithm (DGA) is often perceived as the most difficult to detect using traditional methods. This paper presents an approach for detecting DGA using frequency analysis of the character distribution and the weighted scores of the domain names. The approach's feasibility is demonstrated using a range of legitimate domains and a number of malicious algorithmically-generated domain names. Findings from this study show that domain names made up of English characters "a-z" achieving a weighted score of < 45 are often associated with DGA. When a weighted score of < 45 is applied to the Alexa one million list of domain names, only 15% of the domain names were treated as non-human generated.*

## KEYWORDS

*Domain Generated Algorithm, malicious domain names, Domain Name Frequency Analysis & malicious DNS*

## 1. INTRODUCTION

Domain names and DNS services are often abused by cyber-criminals to provide them with an efficient and reliable communication link for malicious activities [34], [47]. Criminal activities involving Advanced Persistent Threats (APT), malware and botnets use DNS service to locate Command and Control (C&C) servers for file transfer and updates [16], [34]. Spammers also rely on DNS service to redirect users to scams and phishing websites [35]. Zhao et al. [47] explains that these cyber-criminal activities are often successful because DNS traffic is usually unfiltered or allowed through a firewall thereby providing a stealthy and undisturbed communication channel for cyber-criminals to operate.

Cyber-criminals in recent times have been designing malware to take advantage of certain Dynamic DNS (DDNS) capabilities such as the ability to change the IP address associated to a domain [28]. Whilst DDNS provides a useful feature for organisations that need to maintain consistent services, because they rely on a dynamic IP range allocated by their Internet Service Provider (ISP) [18]. Arntz [5] explains that, cyber-criminals exploit this feature to increase the survivability of their C&C server. Zhao et al. also state that, DDNS provide the capability for cyber-criminals to maintain persistent presence on a victim's machine once it has been compromised as they can easily change their IP and domain information [47]. Stevanovic et al. [35] calls DDNS, "agile DNS" and argue that, this feature poses a serious challenge to internet

security. Agile DNS uses dynamic hosting strategies in which domain names and IP addresses associated with a particular service change over time. Well-known DDNS capabilities often exploited by cyber-criminals include: Flux services such as IP-Flux and Domain Flux and the use of Domain Generated Algorithm (DGA) [29]. Amongst the DDNS techniques, DGA is noted as the most elusive and difficult to detect using traditional detection strategies such as signature based solutions.

A number of detection approaches and filtering techniques have been proposed by various scholars and researchers for detecting DGA and flux services. However, criminals continue to adopt DGA because they see it as easy to implement yet difficult to block [5]. Cyber-criminals also employ DGA because it ensures that their C&C network can elude security researchers and law enforcement [14]. Moreover, malware that uses static domain or IP address can easily be taken down or blocked, thereby hindering the operations of the criminal [34].

This piece of work proposes an approach to analysing domain names using frequency analysis of the distribution of letters to ascertain whether the domains were algorithmically-generated or human generated. This work does not seek to compete with the existing tools for detecting DGA but rather it seeks to complement existing works [1], [2], [3], [7], [18],[27],[44]. This paper expands upon previous work related to the detection of algorithmically-generated domain names [15], and seeks to answer the following research question:

1. With what certainty can a letter frequency distribution be used to identify algorithmically-generated domain names and differentiate it from human (legitimate) domains?
2. What are the limitations of systems that have the ability to identify DGA?

## **2. BACKGROUND AND RELATED WORK**

### **2.1. Abuse of Domain Names**

Although there is no single definition of what constitutes domain abuse [41], domain names registered for phishing, malware, botnets and those used for spamming falls into what may be classed as abuse [45]. Often these sorts of activities are recognised in most countries as illegal. Zhao et al. [47] point out that the flexibility of domain names allows cyber-criminals to easily move the IP address of their malware C&C servers using a range of techniques to avoid detection. Zhao et al. [47] also explains that the flexibility of DNS allows cyber-criminals to hide the source of their attack behind proxy servers. To ensure that they remain in control and to maintain persistent access to a compromised device, criminals generally implement some sort of C&C architecture to send and receive information from devices they control.

### **2.2. Malware Command and Control**

Cyber-criminals like to maintain control of devices and hosts they compromise for long term benefits such as financial gain [16]. To achieve this objective, they usually plant some form of a backdoor or create a C&C channel to allow them to re-enter the system at will [26]. C&C activity is initiated when an attacker compromises an end-user device and transforms that device into a bot that listens out for instructions issued by the botmaster [8], [26]. Norton defines a bot as “a type of malware that allows an attacker to take control over an affected computer” [23]. This kind of activity is evident in many cyber-crimes and it is well documented by various writers [16], [33], [47]. Most system hacking methodologies present the implementation of C&C or planting backdoors as a key component of a structured attack [25].





Figure 1: Cyber Kill Chain by Lockheed Martin

Figure 1 above shows an easy to understand model depicting how a cyber-criminal implements C&C in attacking their victim in what is typically known as the Lockheed Martin Cyber Kill Chain [21]. Cyber criminals typically start their attack process with the **recon** (reconnaissance). Under recon the attacker gathers information about the target. Having gathered enough information about the target, the attacker then moves to the next stage of their attack process which involves **weaponisation**. During this phase, the attacker will identify a tool or a piece of malware they intend to use against the target. The attacker will then **deliver** that piece of malware, which they will then use to **exploit** their victim to allow them long term unrestricted access. In order for the attacker to return they will often **install** some form of a backdoor program. Unrestricted access is often achieved using **C&C**. The final stage is the **actions on objectives** which involves financial gain, theft or sabotage.

Although there are several types of C&C communication channels, for example, using of protocols such as Hypertext Transfer Protocol (HTTP), Internet Relay Chat (IRC) and many more [26], traditionally, cyber-criminals tend to rely on a **centralised topology** to control their victims by hardcoding their IP addresses or domain names in their malware binaries to allow persistent access between a server they control and a compromised device [16], [30]. Often, the cyber-criminals include one or more static domains or IP addresses in their source code as observed with the WannaCry ransomware [9], [40].

However, studies have suggested that there are were a number of problems with this approach [2], [16], [47]. Firstly, the hardcoded IP address can be identified by a security researcher reverse engineering the malware, allowing mitigating action to be taken that may involve blacklisting the IP address [16]. Secondly, if the C&C server goes down or the IP address is detected, the nefarious activity can be easily identified and blocked by a security administrator which means their compromised device will be out of the attacker's control [13],[47].

To overcome some of these challenges and to maximise the availability of C&C servers, cyber-criminals have resorted to using a range of methods that avoid detection of their criminal activities yet providing them with high availability and resilience [33]. Chen et al. [13] suggest that the Peer-to-Peer (P2P) botnets infrastructure was one of the first architectural structures that criminals used for overcoming the limitations of the centralised C&C network approach. This is also known as a **decentralised topology**. The Nugache and storm bot utilised the P2P architecture and more recently Waledac and Zeus [2], [16]. The advantage of P2P is that the majority of the bots do not communicate with the centralised machine, hence the IP address is not easily detected to be taken down [13]. However, P2P has some downsides and limitations. They are usually difficult to implement and maintain due to their setup architecture [16]. This has led to the introduction of a **locomotive topology** in which the C&C entities changes over time.

In an effort to combine the simplicity of C&C and the robustness of the P2P infrastructure, cyber-criminals now employ more dynamic strategies to maintain their C&C servers [2], [28]. These new strategies involve generating a large number of domain names using a variety of jurisdictions and service providers for C&C servers [33]. It also involves rapid changes of these domain names through the implementation of Domain Generation Algorithm (DGA). Yadav et al. [44] also points out that spammers also try to avoid detection by systems that rely on regular expression by using DGA.

The objective here is to evade detection or to elude a security engineer's attempt to easily implement a mitigation strategy. Kwon et al. [18] points out that the most common agile DNS implementations associated with malicious activities include: Domain-flux, IP-flux and the use of DGA with some fundamental differences. Zhao et al. [47], however, highlights some similarities such as the "short-life" between Flux services and DGA.

### 2.3. Domain-flux and IP-flux

Berger et al. [17] mentions that the operation of a malicious flux service is similar to a content-delivery network (CDN) service and argues that the theory behind CDN and malicious flux is the same. Zhao et al. [47] explains that whilst CDN is used for facilitating efficient delivery of web-servers content and services, malicious flux is a DNS technique used by criminals to facilitate C&C activity. CDN comprises of a large number of legitimate servers, whereas malicious flux consists of a group of malicious networks. Both Domain-flux and IP-flux involves the continual changing and re-assignment of multiple Fully Qualified Domain Names (FQDNs) to one or more IP addresses to avoid being taken down easily [18], [30]. However, there are some fundamental differences between these two terminologies which are explained below [42]. Examples of some historical malware that employed domain-fluxing are: Cycbot and Murofet [33]. The Zbot and Kazy are also well-known malware that utilised IP-flux or Fast-Flux in their operations [20].

In order to understand domain-flux, the concept of IP-flux is first presented. Historical use of a C&C server by botnets relied only on a single IP address, this meant that once the IP is discovered it was blacklisted and taken down, thereby disabling the communication channel [16]. To get around this problem, criminals began buying a set of IP addresses to ensure that if one IP address is identified, an alternative IP will be introduced to allow the communication to continue, thus implementing IP-flux as shown in Figure 2. However, given that most botnet and malware operating under C&C use DNS to resolve IP addresses to domain names, connecting multiple IP address to a single domain name also results in a single point of failure for the criminal as the malware can be blocked at DNS level [8]. This has led to criminals finding an alternative approach resulting in the development of Domain-flux.

Domain-fluxing uses a DGA to generate many domain names and then attempts to communicate with a small section of the generated domains [37]. Bilge et al. [8] explains that the use of domain names gives the attackers the flexibility and redundancy of migrating their servers with ease. If the malware cannot find the C&C server at its previous domain, it searches the list of domains generated until it finds one that works [16]. Other flux services for example Fast-Flux involve the rapid changes of DNS "A" records. Likewise, there is Double-Flux, in which both the DNS "A" record and the "NS" records are changed rapidly to make take-down much more difficult, yet allowing the cyber-criminal control of their C&C server.

### 2.4. Domain Generation Algorithm (DGA)

According to [27], the first known use of DGA for malicious purposes goes back to the Salinity malware observed in February 2006. Plohmman suggests that Salinity was followed by Torpig in July 2007 and later by the Kraken malware [27]. However, it was not until April 2008 that DGA became publically known when Kraken, Conficker and Szrbi were released [5]. DGA periodically generates algorithmically domain names in order to contact the C&C server. A compromised device will generate a large set of random domain names (Figure 5) and queries each domain until one resolves to a C&C server [47].

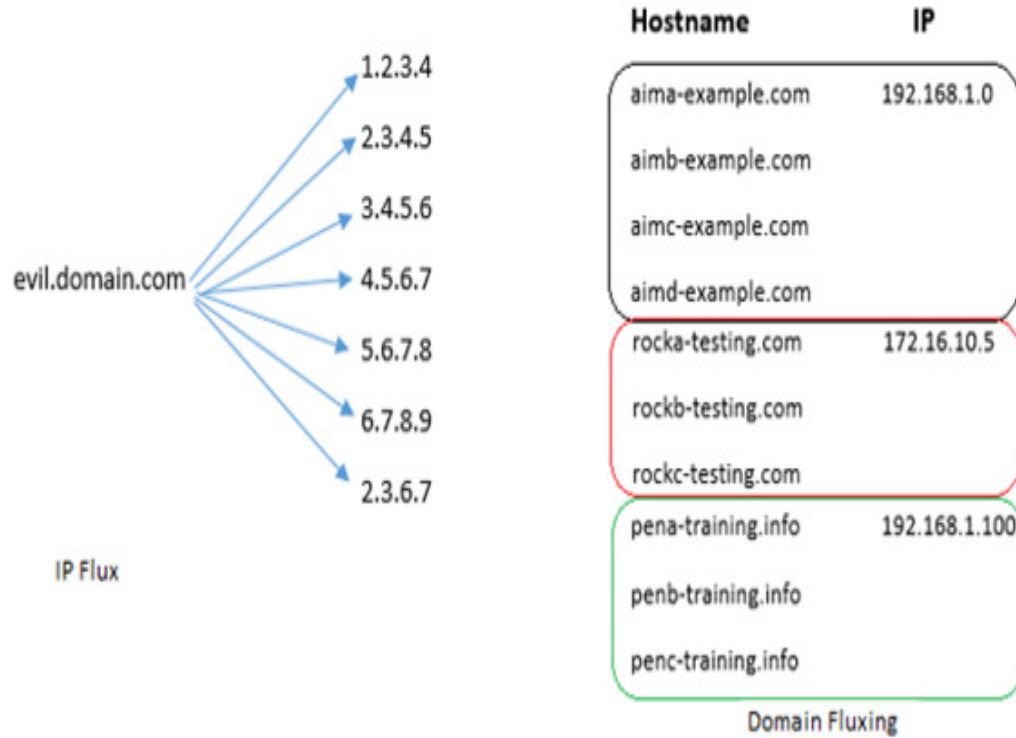


Figure 2: An example of IP-Flux and Domain-Flux

DGA is usually fed with a random seed which consist of numeric constants or a phrase, a time-based element, for example current date and time to dynamically generate often large and multiple FQDNs at run-time [28], [34]. The large number of domain names generated provides great agility and ensures that even if one or more of the domain names are eventually taken down or blacklisted, the compromised device will ultimately get the IP address of the re-allocated C&C server [33]. Antonakakis et al. [2] highlight that a compromised device running DGA will attempt to contact all the generated domains until one of the domains resolves to the IP address of the botmaster's C&C server. From the attacker's perspective, this technique is more advantageous because unless a security researcher can reverse engineer the algorithm the attacker will continue to be in control [47]. Antonakakis et al. [2] also suggest that even when one or two of the C&C servers are taken down, the criminals will simply have to register another domain and the bot will eventually get the IP address of the relocated C&C server. The difficulty is also increased by consistently changing the domain in use. Furthermore, the criminals do not have to include a hard-coded domain within the malware code [28].

Whilst there are some legitimate uses of randomly generated domain names, for example, Plohmman et al. [28] points out that early DGA were used as a backup communication mechanism. Also, Google Chrome queries three random looking domain names when started to act as a DNS check and to verify if Non-Existence Domain (thereafter NXDomain) rewriting is enabled [2], [10]. Ruan et al. [32] argues that in most cases the presence of DGA within network traffic can often signify illegal behaviour of some kind. In fact, Zhao et al. [47] points out that some malware such as Poison Ivy, and Gh0st instructs the attacker to create domains dynamically for locating C&C servers. Antonakakis et al. [2] also points out that Torpig, Srizbi, Bobax, Conficker-A/B and C are all malware that have employed DGA. For example, at its peak the Conficker-C worm, generated almost 50,000 domains per day using DGA [2], [44], [47]. Yet out of the 50,000 domain names generated, Conficker-C only queried roughly 500 of these domains per day. These domains were distributed across 110 Top Level Domain (TLD) [44]. Similarly,

both [36] and [44] highlighted how the Torpig botnet makes use of DGA and utilised the Twitter trend API (Application Programming Interface) as a seed to their algorithm to make detection and take-down difficult for security professionals.

## 2.5. Taxonomy of DGA

DGA differ widely in the way the algorithms are generated and seeded [28], [44]. [11] listed the main types of DGA as: static DGA, Date-based DGA, seed-based DGA and a combination of Date and Seed. Static DGA generates the same domain every time whereas the Date-based method uses the current date as an input for the algorithm to generate domains and the Seed-based DGA utilises hardcoded seed as input for their algorithm. Plohmman et al. [28] on the other hand proposes a taxonomy based on the characteristics of the seed source and identifies four different generation scheme. These include Arithmetic-based DGAs which compute a sequence of characters that have ASCII values constituting the alphabet of the DGA; a Hash-based DGA which relies on hexdigest representation of a hash to generate an algorithmically-generated domain; a wordlist-based DGA which concatenates a string of words and finally there is the permutation-based DGAs. Plohmman et al. [28] explains that permutation-based DGA generates domain names using a permutation of an initial domain name.

Barabosch et al. [6] also highlight four different DGA types shown and points out that, time and causality are the two main possible parameters for any DGA. The first class of DGA family is the deterministic and time independent DGA (**TID-DGA**) and argues that this type of DGA generates the same set of domain names every time they are executed because its use of a static seed in its algorithm. An example of malware that uses TID is Kraken [6]. The second category of DGA family is the time dependent and deterministic DGA (**TDD-DGA**). The seed used in TDD-DGA changes at a regular interval. However, precomputation of the domain names are still easy because it uses a deterministic algorithm. An example of DGA that used TDD-DGA was the Conficker worm. The next type of DGA according to [6] is the non-deterministic and time dependent DGA (**TDN-DGA**). Under TDN-DGA, the seed cannot be anticipated; hence precomputation is not possible. An example of malware that uses TDN was Torpig. Torpig uses the popular trending topics on Twitter as a seed [2]. The final category is the time independent and non-deterministic DGA (**TIN-DGA**). Barabosch et al. [6] suggest that malware that rely on TIN-DGAs have not been seen in the wild yet and argues that this class of DGA might work for small domain name, however, the chances of getting in touch with a C&C server, decreases with the increase of the domain name length.

## 2.6. Previous and related work

Various techniques have been proposed for the detection of Dynamic DNS domain names in DNS traffic. Yadav et al. [44] proposed a methodology for detecting “domain-fluxing and IP-fluxing” in DNS traffic by analysing patterns associated with domain names that are generated algorithmically. They observed that compromised hosts using Domain-fluxing and DGA often generate domain names that exhibit characteristics that are vastly different from legitimate domain names. Their approach relied on signal detection theory, statistical learning strategies and measures such as Kullback-Leibler divergence, Jaccard index and Levenshtein distance to detect algorithmically-generated domain names. They also computed the distribution of alphanumeric characters as well as bi-grams in all domains that are mapped to the same set of IP addresses. Although they suggest that their methodology was successful in detecting malicious domains associated with malwares including Conficker, they mention that when their technique is applied to large data sets of test data their systems efficacy decreases significantly resulting in high false positive rate. For example, the outcome of 50 test words at 100% detection rate produced 50% false positives using a Jaccard index.

Antonakakis et al. [1], also attempts to classify domain names (as malicious or legitimate) by dynamically assigning them a reputation score using a system known as Notos. Notos assigns a low reputation score to a domain suspected of being involved in a malicious activity such as phishing and spam campaigns and a high reputation score to those used for a legitimate purpose. Using a classification and a clustering algorithm, they were able to dynamically blacklist malicious domains, by modelling legitimate and malicious domains, thereby mitigating cyber-attack effectively. The findings of their work had true positive rate of 98%. However, Notos has the limitation of not being able to assign reputation scores to domain names with very little historic information, a key feature of most DGA.

Doyle uses frequency analysis as an approach to detect pseudo-random domain names [15]. His work reports that the pseudo-random generated domain were much more uniform and linear than real-world data sets. However, there were a number of fundamental limitations that makes generalisation difficult. Firstly, the data sample was limited to domains that use .net, .com and .org as their TLD. The problem here is that DGA are not limited to those TLDs. Further, the malicious sample was limited to the Conficker worm. Given that all algorithms are not the same, additional tests will be needed to test whether randomly generated domain names using a variety of algorithms will results in distributions of letters that are vastly different from the letter frequency distribution of human (legitimate) domains.

Bilge et al. suggests a passive DNS analysis technique and detection system known as EXPOSURE that is useful for detecting domain names that are involved in malicious activities [8]. EXPOSURE focuses on extracting certain behavioural features from DNS such as time-based features, for example life-span of the domain (short-life), TTL value and features such as the percentage of numerical characters in a DNS domain. Using these they were able to automatically identify a wide variety of malicious domains. A limitation to their system is that an attacker could adapt their attack to evade features EXPOSURE has been trained to detect. For example an attacker could attempt to change TTL values to evade EXPOSURE. Antonakakis et al. also proposes another passive DNS analysis tool known as Kopis that monitors DNS traffic at the upper levels of the DNS hierarchy [1]. Whilst Kopis is able to detect malicious domain with minimal false positives, it needs a certain mandatory time period to analyse the traffic before results are presented. Hence, DGA bots that operate within a smaller epoch will be inactive by the time Kopis provides its detection results; a limitation of their system.

Antonakakis et al. proposes *Pleiades* that utilises a clustering and classification algorithm to analyse DNS traffic and identify DGA domain names [2],[3]. *Pleiades* relies on domain name requests that result in Name Error Responses (NXDOMAIN) or unsuccessful domain name resolution responses to identify DGA-bots. *Pleiades* searches a cluster of NXDomains that have similar syntactic features and are queried by multiple potentially compromised devices during a given epoch and uses a statistical learning strategy to build a model of the DGA. However, *Pleiades* is unable to learn the exact DGA thereby generating a certain number of false positives and false negatives. Despite these limitations, when applied to real-world DNS traffic obtained from an ISP, *Pleiades* was still able to identify domain names generated using DGA.

Zhao et al. combine signature and anomaly-based solutions to design a detection system that can be placed at the edge of a network to detect DNS traffic that shows signs of malware infection based on 14 certain characteristics extracted through big data analysis of malware associated with DNS activities [14]. Some of the features extracted are similar to those identified by [2] however, their work focuses on detecting Advanced Persistent Threats (APTs), malware and C&C activities within DNS traffic. They highlight certain characteristics of a domain name that can identify potential signs of malicious domains in use. A limitation of their work is that bots that use IP addresses to directly locate the C&C server rather than domains are not detected.

Sharifnya & Abadi [33] proposes the DFBotkiller system that builds on earlier works by other researchers [1], [2], [44]. They argue that previous work does not effectively consider the history of activities on the monitored network; potentially resulting in a high proportion of false positives. To address this problem, they proposed a system that incorporates the history of domain groups linked to malicious activities and suspicious domain failures to automatically assign a reputation score. This approach combines certain characteristics associated with DGA, such as generating a large number of NXDomains and the alphanumeric distribution of Algorithmically-generated domain names, to develop a reputation system that assigns a high negative reputation score to hosts involved in suspicious domain based activities and a low negative score to legitimate domains.

More recent studies by [28] and [43] also present an alternative way of detecting DGA. Plohmann et al. [28] uses a bottom-up approach that relies on reverse engineering to analyse 43 DGA based malware. They identified 253 seeds used in previous and recent DGAs through analysis of domain names to build a ground truth about DGA with no false positive. They built a web-service known as DGArchive to ascertain whether a queried domain originated from a DGA. Wang et al. [43] proposes a DGA-Based detection system known as DBob that monitors only DNS traffic as opposed to previous works that tend to monitor all traffic. DBob, in comparison with previous systems provides a much more efficient way of detecting DGA-based activity as it does not rely on prior training of the system. A limitation of DBob is that devices are clustered based on the similarities of their query behaviours during a detection time window. This means that their system is unable to detect a compromised host if it has never been attacked before or attempted to connect to a C&C server.

## 2.7. Limitations of Machine Learning and Signature Based Systems

A wide variety of complementary techniques exist to detect algorithmically-generated domain names however, these techniques are overly reliant on machine learning and complex computational measures. Whilst machine learning is useful in predicting the possible outcome of an event based on previously learned data sets (training data), Wang et al. highlight that systems that rely on prior training can be evaded by DGA bots if the underlying algorithm or the learned features change [43]. A change in the learned behaviour can easily evade a system based on machine learning [4]. Jadav et al. [44] also points out the resource intensive nature of machine learning. Similarly, when it comes to rule-based systems and Intrusion Detection Systems (IDS) that use signatures, Ruan et al. explains that these systems are insufficient as a strategy as they are unable to respond to the dynamic nature of DGA [13]. All these systems suffer from false positives, making it important for further studies in strategies for detecting malicious DGA.

## 2.8. Frequency Analysis of Letters in a domain name

An alternative strategy to machine learning, reverse engineering of DGA bots and signature-based solution is the use of frequency analysis techniques to examine the character distribution of domain names. Frequency analysis is deeply rooted in descriptive statistics and relates to the number of times an event occurs. As a technique, it is also used in cryptanalysis, to understand how often a character or a group of characters (letters) appears in a cipher-text in order to decrypt and understand a secret text or ciphers for which the key is unknown [38]. This strategy relies on the fact that in any language, certain letters occur with varying frequencies. Whilst the focus of this work is to understand the distribution of alphanumeric characters within a domain name, an appreciation of key measures used within frequency analysis is also important.

[2], [15], [44] includes frequency analysis as part of their strategy to detect DGA albeit they differ in how they employ this technique. *Pleiades* which was developed by Antonakakis et al. uses the

character frequency distribution during the DGA discovery and clustering phase [2]. Yadav et al. for example uses a computation of the bigram of alphanumeric characters (two consecutive characters) within the domain and not the domain or the FQDN [44]. Although [15] attempts to use frequency analysis as a sole technique for the detection of DGA he points the high volume of false positive when using this strategy. However, false positive tend to be present in nearly all the currently available detection strategies. Given the small data sample used by Doyle for his baseline, it is possible that a larger data sample may achieve better results and reduce the level of false positive. This is something that can be investigated in future work.

Doyle [15] compares a small sample of legitimate domains against a well-known English frequency table and proposes a baseline for identifying DGA. His approach is based on a concept well explained by the Unicode Consortium. The Unicode Consortium explains that Domain Names were originally designed to support only the American Standard Code for Information Interchange (ASCII) [39]. They highlight that even non-ASCII Unicode characters are transposed into a special sequence of ASCII characters. This means that as far as DNS system is concerned, all domain names are just ASCII. ASCII as a character encoding standard denotes English characters as numbers with each letter being assigned a number from 0 to 127. Observations from previous studies suggest that the alphanumeric or character distribution of letters in Algorithmically-generated domain names are vastly different from pronounceable words. The fact that domain names are translated into ASCII makes frequency analysis a viable technique for detecting DGA.

Novig [24] points out that although the English language has 26 alphabet letters they do not appear equal amounts of time in an English text. Some letters are used more frequently than others and argues that for instance, the letter “Z” appears less frequently in words than the letter A or E [31]. By using frequency analysis, of letters or a group of letters, the distribution of the characters can be analysed and compared to that of English words to differentiate between legitimate and DGA domains. Figure 3 and Figure 4 below shows the frequency table used.

Letter	Frequency	Letter	Frequency
e	12.7020%	m	2.4060%
t	9.0560%	w	2.3600%
a	8.1670%	f	2.2280%
o	7.5070%	g	2.0150%
i	6.9660%	y	1.9740%
n	6.7490%	p	1.9290%
s	6.3270%	b	1.4920%
h	6.0940%	v	0.9780%
r	5.9870%	k	0.7720%
d	4.2530%	j	0.1530%
l	4.0250%	x	0.1500%
c	2.7820%	q	0.0950%
u	2.7580%	z	0.0740%

Figure 3: Relative Frequency of Letters in the English Language.  
<http://en.algoritmy.net/article/40379/Letter-frequency-English>

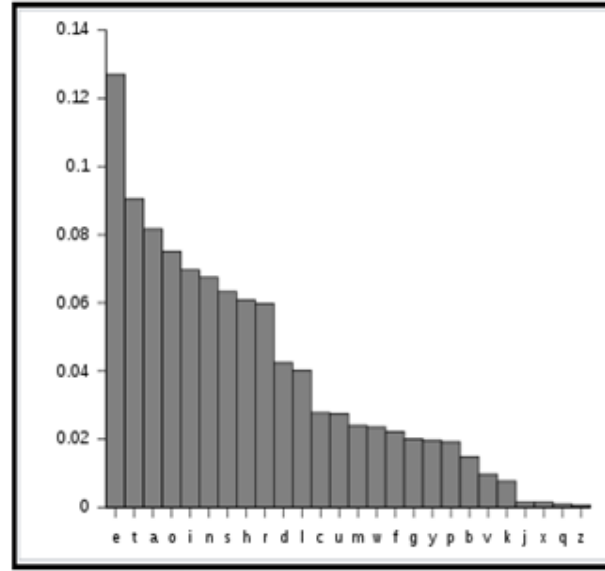


Figure 4: Relative Frequencies of Letters in English text <http://en.algorithmmy.net/article/40379/Letter-frequency-English>

### 3. METHODOLOGY

#### 3.1. Research Approach

The motivation for this work and the approach used here is based on observations by previous studies in this area such as Yadav et al., (2010) and Zhao et al., (2015) that suggests that algorithmically-generated domains do not use pronounceable or real words and hence have a distribution of characters that will be vastly different from legitimate domains. *This hypothesis will be tested using frequency analysis of the distribution of letters in the domain name and if it is true, a schema will be developed that relies on the total weighted score of a domain name to predict potentially suspicious Algorithmically-generated domain names.* Reverse engineering of DGA algorithm will not be covered in this work as it is resource and time intensive [44], moreover reverse engineering of DGA has been previously covered by Plohmman et al., (2016). Classification of different types of DGA and the identification of C&C servers fall outside the remit of this work.

#### 3.2. Selection of Study Samples

The malicious dataset samples were selected based on a convenience samples. However, a simple random technique will be applied to both datasets to select a subset. An observation from the material reviews is that, most DGA tends to have string of characters of length greater than 8. Conficker, for example, which was rated among the top 5 DGA-based crimeware was originally designed to use random strings of between 8-11 characters [14], [29]. According to [17] detecting randomness in words or letters with six (6) characters is often a very difficult task even for humans. This argument is supported by [22] who assert that even humans will often struggle to detect randomness in a string of characters that are less than eight (8) characters in length. With this in mind, a decision was made to select a sample data from the available data set with the characteristics described above.



### 3.3. Data Collection

Both primary and secondary data sets were used in this work. The secondary data was taken from well-known publically available third-party databases to ensure that experimental activity and testing reflected real-life usage. In addition to the secondary data, a prototype DGA that uses only the English alphabetic characters “a-z” was also used as primary data. The reason for including the customised DGA is that, it would allow for testing of a DGA that uses only the English 26 alphabetic characters to examine how the character distribution will compare. There are different types of DGA and it is impossible to predict the kind of algorithm being utilised by the cyber-criminal hence the prototype allows for some form of testing of a previously unknown DGA [11]. The legitimate domains are obtained from Alexa. Alexa has acted as the source of data for a number of other works in this area [8], [19], [18], [33]. The malicious or algorithmically-generated domain names were sourced from third parties such as bambenekconsulting.com.

### 3.4. Frequency Analysis of the domain names

A customised script was applied to the domain names (excluding protocols and TLD’s) to generate letter frequency distribution. The character distribution of legitimate domains was compared to a standard frequency distribution table of English text to examine their relationship. Doyle [15] observed a close similarity between legitimate domains and a well-known English frequency table, albeit his sample of legitimate domains was relatively small in comparison to the sample used in this study. Doyle [15], propose a weighted scoring technique that assigns scores to the characters and uses this information as the basis to detect a DGA. Figure 5 below shows the formula for assigning the weighted scores. The weighted score approach described above will be used to produce a cumulative frequency plot for further statistical analysis of the domains.

### 3.5. Weighted score-based analysis

A weighted score-based analysis was used in this work to assign scores to the domain names. However, there are differences between this work and the work conducted by [15]. Firstly, the upper limit of his domain weighted score will differ from this study because of the size of his good data which was used as the reference point. Secondly, his analysis of the domains were limited to second level domains. This work uses the entire domain excluding TLDs and protocols. Thirdly, the malicious dataset used in [15] was limited to a single malware family (Conficker). This does not give a good indication of how well his system will perform when used for the detection of other DGA based malware domains.

$$w = \frac{\sum_{i=0}^n x_i}{n} \times 1000$$

*Where:*  
*w = weighting*  
*n = num. letters in domain*  
*x = frequency lookup of letter i*

Figure 5: Weighted Score Formula

## 4. EXPERIMENTAL ACTIVITIES

### 4.1. Analysis of the Data Sample

The non-malicious dataset (legitimate domain) used in this work was sourced from Seobook.com and contains the Alexa list of one million domains. The malicious datasets (DGA samples) were taken from bambenekconsulting.com and comprises of 5 days of DGA feeds. The DGA samples

differ in the way in which the algorithms were seeded. Differentiating these samples of malware is important because as pointed out by [44], different malware employing DGA are seeded differently. As an example, Conficker generated 250 domains every three hours using the current date and time at UTC as the seed. Torpig's seeding was based on the most popular trending topics on twitter [44]. The DGA samples obtained include: Cryptolocker, Kraken, Tinba, P2P Gameover, Zeus, Shifu, Banjori, Pykspa, Dyre, Ramnit, Cryptowall, Ranbyus, Simda, Murofet, Qakbot and Necurs. All TLDs and protocols were removed leaving just the domain names. Stripping off the TLD was crucial because there are over a thousand TLD's, any of which can be used by criminals [44]. It is difficult to predict which TLD a criminal will use. It was also felt that adding the TLD would not give an accurate representation of letters used in the generation of the domains.

The legitimate domains were used as the baseline or reference point of what a "good domain" should look like and domain names that differed vastly (based on measure of dispersion) from this baseline are treated as suspicious. In other words, the distribution table of well-known (legitimate domains) acted as the reference table for all good domains. Three groups of malicious samples; Ramnit, Tinba and Qakbot each containing 40,000 domains were selected for the initial experimental activity. The reason behind selecting 40,000 is that there is already a frequency table for 40,000 English words as observed in the literature review, hence where a comparison is required, that frequency table can be used. A customised script was applied to the domain names to create a frequency table of the character distribution.

Table 1: Character Frequency distribution of a random sample of 40,000 DGA domain names against the Alexa one million domains.

Relative Frequency of letters in English text.	English Text	Alexa 1 million	40,000 English Words	40,000 Random DGA
E	0.12702	0.095	0.1202	0.0429
T	0.09056	0.061	0.091	0.0375
A	0.08167	0.0921	0.0812	0.0294
O	0.07507	0.074	0.0768	0.0392
I	0.06966	0.0722	0.0731	0.0426
N	0.06749	0.06	0.0695	0.0396
S	0.06327	0.0645	0.0628	0.0375
H	0.06094	0.0249	0.0592	0.0428
R	0.05987	0.0639	0.0602	0.0376
D	0.04253	0.0322	0.0432	0.0425
L	0.04025	0.0472	0.0398	0.0391
C	0.02782	0.0379	0.0271	0.0428
U	0.02758	0.0333	0.0288	0.0375
M	0.02406	0.0338	0.0261	0.0392
W	0.0236	0.012	0.0209	0.0374
F	0.02228	0.0171	0.023	0.0421
G	0.02015	0.0243	0.0203	0.0421
Y	0.01974	0.0163	0.0211	0.0369

P	0.01929	0.03	0.0182	0.0379
B	0.01492	0.0238	0.0149	0.0434
V	0.00978	0.0136	0.0111	0.0374
K	0.00772	0.0188	0.0069	0.0396
J	0.00153	0.0055	0.001	0.0425
X	0.0015	0.0065	0.0017	0.0372
Q	0.00095	0.0021	0.0011	0.0375
Z	0.00074	0.0065	0.0007	0.0154

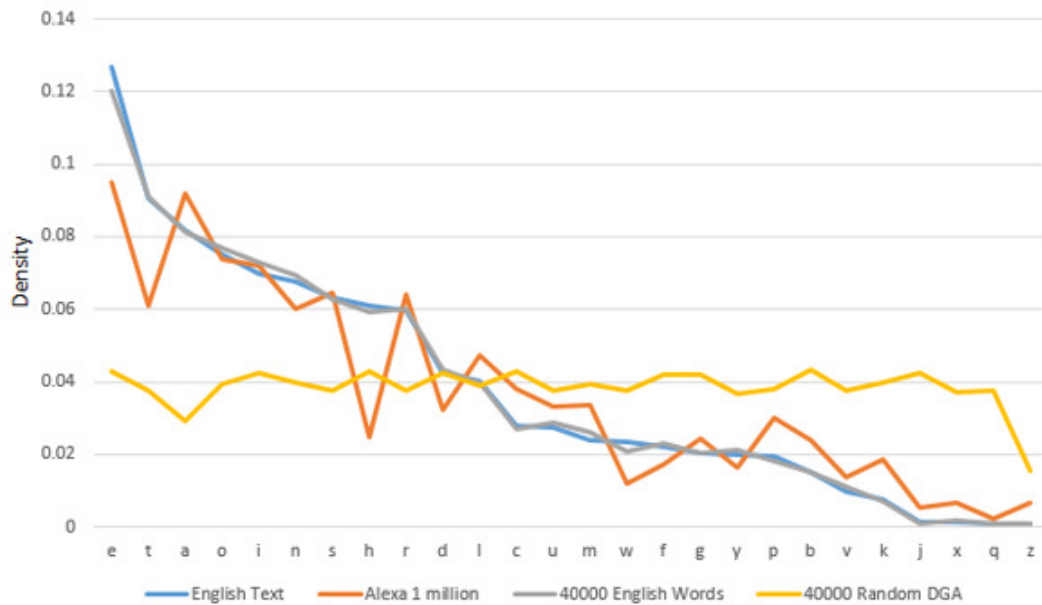


Figure 6: A line chart showing how a random sample of 40,000 DGA domain names compare to English words and text frequency table.

For the purposes of this work, the list of domains from Alexa is assumed as the “correct” frequency distribution for domain names and will therefore be used as the baseline. Domains that differ vastly from this baseline are treated as suspicious.

## 4.2 Findings

The character frequency distribution of the domain names from Alexa does show a close similarity with the English letter frequency table, albeit there were one or two characters such as “t” and “h” that had a wide variation as observed in Table 1 and Figure 6. In order words the frequencies are different from English text. The similarities can be linked to the fact that many “domain names follow English nouns, verbs, adjectives or a combination of these” [15].

By applying the above formula in Figure 5, the frequency of characters in the domain name were added together and divided by the number of characters in the domain. The outcome was then multiplied by 1000 to facilitate the ease of calculation. Each domain that was calculated was expected to fall within a range of:

$$0 < w < 95$$

The limit of 95 is obtained by multiplying the highest frequency letter “e” from the good domain (Alexa). This limit is roughly 10% less than the upper limit used by [15]. All the scores for the domain should fall within the range; when using the frequency table of the good domains taken from the Alexa top 1 million. For instance by applying this formula (using a customised script), the weighted score for the domain *dailymail.co.uk* and *americanexpress.com* can be computed.

When the TLD is excluded, “dailymail” will get a score of 56.1444 whereas “americanexpress” will get a score of 64.4267.

The calculations for “dailymail” is shown below:

$$(0.0322_d + 0.0921_a + 0.0722_i + 0.0472_l + 0.0163_y + 0.0338_m + 0.0921_a + 0.0722_i + 0.0472_l) / 9 \times 1000 = 56.14444$$

Likewise the formula when applied to “americanexpress” achieves a score of 64.4267.

$$(0.0921_a + 0.0338_m + 0.095_e + 0.0639_r + 0.0722_i + 0.0379_c + 0.0921_a + 0.06_n + 0.095_e + 0.0065_x + 0.03_p + 0.0639_r + 0.095_e + 0.0645_s + 0.0645_s) / 15 \times 1000 = 64.4267$$

The weighted scores of legitimate domains could be compared against the DGA domains. Doyle (2010) used the cumulative weighted scores of the domain names to analyse 3381 samples of Conficker DGA. Although, he had some interesting results including being able to block 87.35% of pseudo-random domains, his reference sample (legitimate data) was not large enough to provide a baseline of what a good domain should be like. In addition, his legitimate domains were limited to domains that use .net, .org, and .com. Furthermore, his study was limited to analysis of Conficker which makes it difficult to generalise his findings. This project uses a larger data sample and assumes the one million list of domains as a good reference table for domain names compared to the 3381 legitimate and malicious DGA used in [15]. The cumulative weighting score for a random sample of legitimate and malicious DGA comprising of Qakbot, Tinba and Ramnit is shown in Figure 7 below:

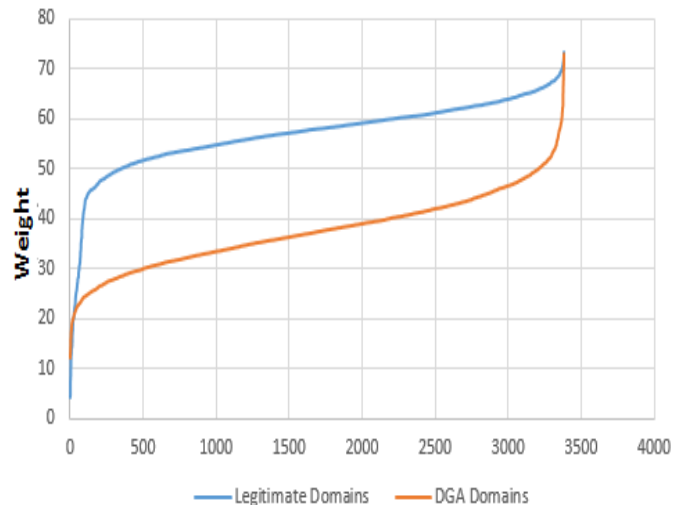


Figure 7: A plotted cumulative weights of malicious domains versus legitimate domain names.

The plotted graph (Figure 7) and the data used for the generation of the graph can be used to carry out some basic blocking of malicious DGA.

**4.2.1. Test 1 - Block 80% of the sample DGA and determine the percentage of false positive.**

80% blocking of the DGA domains would require a score of  $< 43.538$  to be blocked. This score when applied to the legitimate domains would result in 3.13% blocking of these domains (over-blocking). The over-blocking in this experiment seemed to achieve a better result than [15]. This could be put down to using a large legitimate sample to create the baseline of the assumed “correct” frequency table of legitimate domains.

**4.2.2. Test 2 - Block 95% of the sample DGA and determine the percentage of false positives.**

This test increased the number of DGA domains that need to be blocked to 95%. A score of  $< 50.39$  would need to be blocked in this instance. Applying this score to the domains would result in over-blocking of the legitimate samples by 11.56%.

**4.2.3. Test 3 - Aim for a 20% false positives.**

This test set a limit on the number of false positives generated from the legitimate domains. 20% false positive rate is too high for real-world application [15]. However, this test still gives a good indication of how the proposed schema will respond under this condition. In this situation a score of  $< 52.95$  is required. This results in 97.5% blocking of the DGA domains.

**4.2.4. Test 4 - Aim for no more than 5% false positives.**

This test set a limit on the number of false positive to 5%. This would require a score of  $< 46.22$  to be blocked. This score achieves a blocking rate of 88.7% of the DGA domains. Given that the information being used is the character distribution of the letters in the domains, blocking 88.7% of the algorithmically-generated domain name with 5% over-blocking of legitimate domain would be reasonable. Although the tests achieved some impressive outcomes with the data set, there are still the issues of false positives. Plohmman et al. points out that malware that uses DGA has different characteristics and features, in terms of how they are seeded and how the algorithm works [28]. With this in mind, additional testing was conducted with a range of DGA. Overall findings shows that in almost all cases the malicious sample when plotted were linear in comparison to non-malicious samples.

**5. DISCUSSION AND IMPLICATION****5.1. Efficiency of the design approach**

Overall, the experimental activities were successful in terms of the ability to detect DGA domains and the results obtained supports findings from previous studies and the hypothesis that states that the distribution of alphanumeric characters in DGA domain names is vastly different from that in legitimate domain names. A top ceiling weighted score of 95 was applied to all domain names. By doing this, no domain name was expected to have a weighted score greater than 95. This is lower than the numerical value of 106 used by [15]. The discrepancies in these figures can be attributed to the large sample size used in this study. As expected none of the domains had a weighted score greater than 95.

In almost all cases, the malicious sample shows a linear distribution when compared to legitimate domain names. For example, Tinba, Pyskpa, Ramnit and Qakbot were all observed to have a linear distribution. Legitimate domains tend to follow a similar pattern to the frequency

distribution of well-known English words and text. This observation can be attributed to the fact that most domain names follow nouns, verbs, adjectives and or a combination of these [15]. It can also be argued that humans will often use a memorable pattern of words whereas machine generated algorithms do not have such limitations.

In comparing this work to other approaches, it is evident that frequency analysis of the characters in the domain name combined with the weighted score approach can dynamically detect Algorithmically-generated domains in a more efficient manner than systems that rely on static signatures albeit there are some limitations. This approach does not have the same limitations as signature based solutions that lack the flexibility to respond to changes in behaviour. The statistical-based approach overcomes the non-dynamic nature of rule-based solutions. Even though this approach is solely dependent on frequency analysis of the letters in the domain names, it achieves a lower false positive when tested against some malicious DGA domain names than the KL and edit distance used by [44]. Yadav et al. reports 50% false positives in their work [44]. Other systems however do have a much lower false positive at the detection of DGA. For example, Kopis proposed by Antonakakis et al. has a much smaller false positive rate [1].

## 5.2. Study Limitations

Despite, some of the good outcomes generated using the proposed approach, there are a number of disadvantages to the weighted based approach. Weights and scores are susceptible to numerical variances due to human bias and other uncertainties [46]. This can be seen in the upper limits that are applied in this work. Whilst the research did not directly manipulate the ceiling of the domain names, there is uncertainty of what that ceiling will be if this test is repeated using a different number of data set.

In addition to the limitations of the weighted score approach highlighted by [46], there are problems with very short domain names that will always result in a much lower score in comparison to longer domain names. For example, bt.com and cnn.com for example will generate a much lower score than a domain such as americanexpress.co.uk or dailymail.co.uk. Also, there is nothing stopping criminals from using algorithms or seeding techniques that generates domains that fall within legitimate English text or letter frequency distribution. The approach presented in this work does not take into consideration domain names that appear to be machine-generated but are used for legitimate purposes. Also, provision was not made for misconfigured applications or domain names resulting from typos. These may be flagged as possible DGA. Future work could focus on addressing these issues.

## 6. CONCLUSION AND IMPLICATION OF THE FINDINGS

Based on the experimental activities, it was observed that domain names achieving a weighted score of  $< 45$  or domains with a weighted score of approximately 20% deviation from the average weighted score of Alexa 1 million are often Algorithmically-generated (non-human generated). The average score of the domains on the Alexa list is 55.19. A legitimate domain (human generated) typically achieves a weighted score  $> 45$  with a top limit of 95 based on earlier calculations. When a weighted score of  $< 45$  is applied to the Alexa one million list of domains, only 15% of domains were treated as non-human generated. This is fairly impressive given that the detection is based purely on weights of the domain. Also, a group of NXDomains at a given epoch achieving a score of  $< 45$  would be deemed as potentially malicious for further investigation. The issue of blocking legitimate traffic means that this approach can have a negative impact in real-life applications if it is used as a sole approach, rather the weighted based approach can be used by DGA detection systems to help access the overall validity of a domain. Future work could investigate how to reduce or eliminate the blocking of legitimate domain by

measuring the randomness of characters in domain names by using measures such as Kolmogorov Complexity (K-complexity) and Hidden Markov Model (HMM).

## ACKNOWLEDGEMENTS

Many thanks also goes to Adam Wedgbury of Airbus Research Group Innovations for the initial idea. We will also like to thank Dr Ebenezer Paintsil, Shane Murnion and Richard Macfarlane for their feedback.

## REFERENCES

- [1] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou II, and D. Dagon, (2011, August) “Detecting Malware Domains at the Upper DNS Hierarchy”. In USENIX security symposium Vol. 11, pp. 1-16, 2011
- [2] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou II, S. Abu-Nimeh, W. Lee, and D. Dagon, “From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware”. In USENIX security symposium Vol. 12, 2012.
- [3] M. Antonakakis, R. Perdisci, N. Vasiloglou, and W. Lee, Detecting and Tracking the Rise of DGA-Based Malware. The magazine of USENIX & SAGE, 37(6), 15-24, 2012
- [4] H. Armstrong, (2015, July 05). Machine that learn in the wild. Available: [https://www.nesta.org.uk/sites/default/files/machines\\_that\\_learn\\_in\\_the\\_wild.pdf](https://www.nesta.org.uk/sites/default/files/machines_that_learn_in_the_wild.pdf)
- [5] P. Arntz, (2016, June 27). Explained: Domain Generating Algorithm. Available: <https://blog.malwarebytes.com/security-world/2016/12/explained-domain-generatingalgorithm/>
- [6] T. Barabosch, A. Wichmann, F. Leder, and E. Gerhards-Padilla, (n.d.). Automatic Extraction of Domain Names Generation Algorithms from Current Malware. Available: [https://net.cs.unibonn.de/fileadmin/user\\_upload/wichmann/Extraction\\_DNGA\\_Malware.pdf](https://net.cs.unibonn.de/fileadmin/user_upload/wichmann/Extraction_DNGA_Malware.pdf)
- [7] A. Berger, A. D’Alconzo, W.N. Gansterer, and A. Pescapé, “Mining agile DNS traffic using graph analysis for cybercrime detection”. Computer Networks, 100, 28-44, 2016
- [8] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, (2011, February). EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis. In Ndss.
- [9] R. Bandom, (2017, May 17) Registering a single web address may have stopped a global malware attack -Finding the kill switch. Available: <https://www.theverge.com/2017/5/13/15635050/wannacry-ransomware-kill-switchprotect-nhs-attack>.
- [10] CERT Polska (2015, May 26). DGA botnet domains: on false alarms in detection. Available: <https://www.cert.pl/en/news/single/dga-botnet-domains-false-alarms-in-detection/>
- [11] A. Chailtyko, and A. Trafimchuk, (2015, July 17). DGA clustering and analysis: mastering modern, evolving threats. Available: <https://www.botconf.eu/wp-content/uploads/2015/12/OK-S01-Alex-Chailtyko-Alex-Trafimchuk-DGA-clustering-and-analysis-mastering-modern-evolvingthreats.pdf>.
- [12] M. Chapple, M. (2017, July 28.). Evaluating and tuning an Intrusion Detection System. Available: <http://searchsecurity.techtarget.com/tip/Evaluating-and-tuning-an-intrusion-detectionsystem>.
- [13] R. Chen, W. Niu, X. Zhang, Z. Zhuo, and F. Lv, “An Effective Conversation-Based Botnet Detection Method”. Mathematical Problems in Engineering, 2017.

- [14] Damballa (2012, July 17). DGAs in the Hands of Cyber-Criminals: Examining the state of the art in malware evasion techniques. Available: [https://www.damballa.com/wpcontent/uploads/2014/02/WP\\_DGAs-in-the-Hands-of-Cyber-Criminals.pdf](https://www.damballa.com/wpcontent/uploads/2014/02/WP_DGAs-in-the-Hands-of-Cyber-Criminals.pdf) Accessed.
- [15] R. Doyle, (2010, June 17). Frequency analysis of second-level domain names and detection of pseudorandom domain generation. Available: [http://ryandoyle.net/assets/papers/Frequency\\_analysis\\_second\\_level\\_domains\\_June\\_2010\\_RDoyle.pdf](http://ryandoyle.net/assets/papers/Frequency_analysis_second_level_domains_June_2010_RDoyle.pdf)
- [16] N. Goodman, A Survey of Advances in Botnet Technologies. arXiv preprint arXiv:1702.01132, 2017
- [17] A. Kololkoltsev, (2015, July 28). Machine learning technique to detect generated domain names. Available: [https://www.youtube.com/watch?v=9wB\\_ovM5C0M](https://www.youtube.com/watch?v=9wB_ovM5C0M).
- [18] J. Kwon, J. Lee, H. Lee, and A. Perrig, "PsyBoG: A scalable botnet detection method for large-scale DNS traffic". Computer Networks, 97, 48-73, 2016
- [19] J. Lee, and H. Lee, "GMAD: Graph-based Malware Activity Detection by DNS traffic analysis". Computer Communications, 49, 33-47, 2014
- [20] D. Mahjoub, (2013, September). "Monitoring a fast flux botnet using recursive and passive DNS: A case study". In eCrime Researchers Summit (eCRS), 2013 (pp. 1-9). IEEE.
- [21] L. Martin, (2014). "Cyber Kill Chain®". Available: [http://cyber.lockheedmartin.com/hubfs/Gaining\\_the\\_Advantage\\_Cyber\\_Kill\\_Chain.pdf](http://cyber.lockheedmartin.com/hubfs/Gaining_the_Advantage_Cyber_Kill_Chain.pdf).
- [22] M. Namazifar, (2015, July 17). Detecting Random strings: A language based approach. Available: <https://www.youtube.com/watch?v=70q5ojxNuv4>.
- [23] Norton (2016, July 17). Bots and Botnets. Available: <https://us.norton.com/botnet/>
- [24] P. Norvig, (2012). English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLCU. Available: <http://norvig.com/mayzner.html> Accessed 02 July 2017
- [25] S.P. Oriyano, CEH v9: Certified Ethical Hacker Version 9 Study Guide. John Wiley & Sons 2016
- [26] V. Oujezsky, T. Horvath, and V. Skorpil, "Botnet C&C Traffic and Flow Lifespans Using Survival Analysis". International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems, 6(1), 38-44, 201
- [27] D. Plohmann, (2015). DGAArchive – A deep dive into domain generating malware. Available: <https://www.botconf.eu/wp-content/uploads/2015/12/OK-P06-Plohmann-DGArchive.pdf>
- [28] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla, (2016). "A Comprehensive Measurement Study of Domain Generating Malware". In 25th USENIX Security Symposium (USENIX Security 16) pp. 263-278, 2016 USENIX Association.
- [29] P. Porras, H. Saidi, and V. Yegneswaran, V. "An analysis of Conficker's logic and rendezvous points". Technical report, SRI International. 2009
- [30] M. Poor, SANS 503: Intrusion Detection in-depth. The SANS institute, 2015
- [31] D. Rodriguez-Clark, (2017). Frequency Analysis: breaking the code . Available: <http://crypto.interactive-maths.com/frequency-analysis-breaking-the-code.html>
- [32] W. Ruan, Y. Liu, and R. Zhao, "Pattern discovery in DNS query traffic". Procedia Computer Science, 17, 80-87, 2013



- [33] R. Sharifnaya, and M. Abadi, DFBotKiller: Domain-flux botnet detection based on the history of group activities and failures in DNS traffic. *Digital Investigation*, 12, 15-26, 2015.
- [34] U. Sternfeld, (2016). Dissecting domain generation algorithm: eight real world DGA Variants. Available: <http://go.cybereason.com/rs/996-YZT-709/images/Cybereason-Lab-Analysis-Dissecting-DGAs-Eight-Real-World-DGA-Variants.pdf>
- [35] M. Stevanovic, J.M. Pedersen, A. D'Alconzo, S. Ruehrup, and A. Berger, "On the ground truth problem of malicious DNS traffic analysis". *Computers & Security*, 55, 142-158, 2015.
- [36] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, and G. Vigna, (2009, November). Your botnet is my botnet: analysis of a botnet takeover. In *Proceedings of the 16th ACM conference on Computer and communications security* (pp. 635-647). ACM.
- [37] A. Almomani "Fast-flux hunter: a system for filtering online fast-flux botnet". *Neural Computing and Application* 29(7), 483-493, 2018
- [38] C. Swenson, *Modern cryptanalysis: techniques for advanced code breaking*. John Wiley & Sons, 2008
- [39] The Unicode Consortium. Internationalized Domain Names (IDN) FAQ. Available: <http://unicode.org/faq/idn.html>. Accessed 07 June 2017.
- [40] US-CERT (2015). Indicators Associated with WannaCry Ransomware. Available: <https://www.us-cert.gov/ncas/alerts/TA17-132A> Accessed 30 May 2017
- [41] P. Vixie, "What DNS is not". *Commun. ACM*, 52(12), 43-47, 2009
- [42] L. Vu Hong, (2012). *DNS Traffic Analysis for Network-based Malware Detection*.
- [43] T.S. Wang, H.T. Lin, W.T. Cheng, and Chen, C. Y. "DBod: Clustering and detecting DGA-based botnets using DNS traffic analysis". *Computers & Security*, 64, 1-15, 2017
- [44] S. Yadav, A.K.K Reddy, A.L. Reddy, and S. Ranjan, (2010, November). Detecting Algorithmically-generated malicious domain names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement* (pp. 48-61). ACM.
- [45] M. Young, M. (2014). Domain name abuse is a 4 letter word. Available: [http://www.circleid.com/posts/20141112\\_domain\\_name\\_abuse\\_is\\_a\\_4\\_letter\\_word/](http://www.circleid.com/posts/20141112_domain_name_abuse_is_a_4_letter_word/)
- [46] J. Yuventi, and S. Weiss, (2013). Probabilistic Consideration Method for Weight/Score-Based Decisions in Systems Engineering-Related Applications.
- [47] G. Zhao, K. Xu, L. Xu, and B. Wu, (2015). "Detecting APT Malware Infections Based on Malicious DNS and Traffic Analysis". *IEEE Access*, 3, 1132-1142, 2015

**AUTHORS**

**Enoch Agyepong** is a Lead Cyber Security Engineer at Airbus. He holds Bachelor's Degree in Computing and Master's Degree in Advanced Security And Digital Forensics from Edinburgh Napier University, United Kingdom



**William J. Buchanan** is a Professor in the School of Computing at Edinburgh Napier University, and a Fellow of the BCS. He was awarded an OBE in the Queen's Birthday awards in June 2017. Bill currently leads the Centre for Distributed Computing, Networks, and Security and The Cyber Academy. He has published over 27 academic books, and over 250 academic research papers, along with several awards for excellence in knowledge transfer, and for teaching. Bill has been included in a list of the "Top 50 Scottish Tech People Who Are Changing The World".



**Kevin Jones** is the Head of Cyber Security Architecture, Innovation and Scouting. He holds PhD in Computer Science and Mathematics from De Montford University, UK



# TWO DISCRETE BINARY VERSIONS OF AFRICAN BUFFALO OPTIMIZATION METAHEURISTIC

Amira GHERBOUDJ

MISC laboratory, NTIC faculty, AbdelhamidMehri University Constantine 2,  
Algeria

## **ABSTRACT**

*African Buffalo Optimization (ABO) is one of the most recent swarms intelligence based metaheuristics. ABO algorithm is inspired by the buffalo's behavior and lifestyle. Unfortunately, the standard ABO algorithm is proposed only for continuous optimization problems. In this paper, the authors propose two discrete binary ABO algorithms to deal with binary optimization problems. In the first version (called SBABO) they use the sigmoid function and probability model to generate binary solutions. In the second version (called LBABO) they use some logical operator to operate the binary solutions. Computational results on two knapsack problems (KP and MKP) instances show the effectiveness of the proposed algorithm and their ability to achieve good and promising solutions.*

## **KEYWORDS**

*Optimization, metaheuristic, swarm intelligence, binary problems, African Buffalo Optimization, knapsack problems*

## **1. INTRODUCTION**

Solving optimization problem is finding a solution of sufficient quality (i.e. optimal or near optimal) from a set of solutions taking into account the constraints imposed and objectives to meet. It is to maximize or minimize one or a set of fitness functions respecting constraints of the treated problem. This research line that sought the attention of several research teams is intrinsically linked to operational research and it uses mathematical and computer tricks.

Various Methods have been proposed to solve optimization problems. They are often classified into two classes: exact methods and approximate methods. The prohibitive cost associated with exact methods has excited researchers to use approximate methods. The investigation in the area of approximate methods gave rise to a subclass of approximate methods called "Metaheuristics". They are general and applicable methods on a wide range of optimization problems.

Metaheuristics based on swarm intelligence have built a very active trend over the last decade. They are generally inspired by the collective behavior of some species in evolving in groups and

solving their problems. These species gather in swarm to build a collective force that allows them to surpass their very limited individual capacities.

The African Buffalo Optimization (ABO) is one of the most recent swarm intelligence based Metaheuristics. It was proposed in 2015, by Julius BeneoluchiOdili et al [1]. ABO is inspired from the behavior of African buffaloes in the vast African forests and savannahs [1]. The recent applications of ABO Metaheuristic for optimization problems have shown its promising effectiveness as it has been proven in [2], [3] and [4]. The original ABO algorithm is a continuous version, which solves only continuous problems. The aim of this paper is to propose two binary versions of ABO algorithm to cope with binary optimization problems. The main difference between the original version of ABO algorithm and the proposed binary versions is that, in the original ABO algorithm, the solution is composed of a set of real numbers. While in the proposed binary versions, the solution is composed of a set of bits.

The remainder of this paper is organized as follows. Section 2 presents an overview of African Buffalo Optimization (ABO) Algorithm. The proposed algorithms (SBABO and LBABO) are presented in section 3. Experimental results are presented and discussed in section 4, and a conclusion and perspectives are provided in the fifth section of this paper.

## 2. AFRICAN BUFFALO OPTIMIZATION

In order to solve complex problems, ideas gleaned from natural mechanisms have been exploited to develop heuristics. Nature inspired optimization algorithms has been extensively investigated during the last decade paving the way for new computing paradigms such as neural networks, evolutionary computing, swarm optimization, etc. The ultimate goal is to develop systems that have the ability to learn incrementally, to be adaptable to their environment and to be tolerant to noise. One of the recent developed bioinspired algorithms is the African Buffalo Optimization algorithm.

The African Buffalo Optimization is inspired from the cooperative and competitive behavior of buffaloes. ABO models the three characteristic behaviors of the African buffaloes that enable their search for pastures. First is their extensive memory capacity. This enables the buffaloes to keep track of their routes. The second attribute of the buffaloes is their cooperative communicative ability whether in good or bad times. The third attribute of the buffaloes is their democratic nature borne out of extreme intelligence. In cases where there are opposing calls by members of the herd, the buffaloes have a way of doing an ‘election’ where the majority decision determines the next line of action [1]. Furthermore, ABO algorithm models the two sounds for communication that buffaloes use to exploit and explore the search space:

- The warning sound “waaa” with which they ask the herd to keep moving because the present location is unfavorable, lacks pasture or is dangerous. This sound encourages the buffaloes to explore the research space.
- The alert sound “maaa” with which they stay on the present location because it holds promise of good grazing pastures and is safe. This sound encourages the buffaloes to exploit the research space

Algorithm1 presents a pseudo algorithm of the ABO method.

**Algorithm 1: ABO Algorithm**

1. Objective function  $f(x) = (x_1, x_2, \dots, x_n)^T$ ;
2. Initialization: randomly place buffaloes to nodes at the solution space;
3. Update the buffaloes fitness values using (1);
4. Update the location of buffalo k ( $bp_{max.k}$  and  $bg_{max}$ ) using (2);
5. Is  $bg_{max}$  updating. Yes, go to 6. No, go to 2;
6. If the stopping criteria is not met, go back to algorithm step 3, else go to step 7;
7. Output the best solution.

The generation of new solutions is done by using equations 1 and 2.

$$m_{.k+1} = m_{.k} + lp_1(bg_{max} - w_{.k}) + lp_2(bp_{max.k} - w_{.k}) \quad (1)$$

$$w_{.k+1} = \frac{w_{.k} + m_{.k}}{-+0.5} \quad (2)$$

Where:

- $w_{.k}$  and  $m_{.k}$  presents the exploration and exploitation moves respectively of the  $k^{th}$  buffalo ( $k=1,2,\dots,N$ );
- $lp_1$  and  $lp_2$  are learning factors;
- $bg_{max}$  is the herd's best fitness;
- $bp_{max.k}$  the individual buffalo's best.

### 3. THE PROPOSED DISCRETE BINARY VERSIONS

Optimization problems can be classed into two main classes: continuous optimization problems and discrete optimization problems. In continuous optimization problems, the solution is presented by a set of real numbers. However, in discrete optimization problems, the solution is presented by a set of integer numbers. Discrete binary optimization problems are a sub-class of the discrete optimization problems class, in which a solution is presented by a set of bits. Many optimization problems can be modelled as a discrete binary search space such as, flowshop scheduling problem [5], job-shop scheduling problem [6], routing problems [7], KP [8] and its variants such as the MKP [9], the quadratic KP [10], the quadratic multiple KP [11] and so one.

The original ABO algorithm operates in continuous search space. It gives a set of real numbers as a solution of the handled problem. However, a binary optimization problem needs a binary solution and the real solutions are not acceptable because they are considered as illegal solutions. In the aim to extend the ABO algorithm to discrete binary areas, we propose in this paper two binary versions of ABO that we called SBABO and LBABO. The main objective of the SBABO and LBABO algorithms is to deal with the binary optimization problems.

### 3.1 SBABO ALGORITHM

In the SBABO algorithm, we introduce a binarization phase of solutions in the core of the original ABO in order to obtain a binary solution for the treated problem. The objective of this phase (i.e. binarization) is to transform a solution  $x_i$  from real area to binary area. To meet this need, we propose to constrain the solution  $x_i$  in the interval  $[0, 1]$  using the Sigmoid Function as follows:

$$S(x_i) = \frac{1}{(1 + e^{-x_i})} \quad (3)$$

Where  $S(x_i)$  is the flipping chance of bit  $x'_i$ . It presents the probability of bit  $x'_i$  takes the value 1. To obtain the binary solution  $x'_i$ , we have to generate a random number from the interval  $[0,1]$  for each dimension  $i$  of the solution  $x$  and compare it with the flipping chance  $S(x_i)$  as mentioned below in equation (4). If the random number is lower than the flipping chance of bit  $x'_i$ ,  $x'_i$  takes the value 1. Otherwise,  $x'_i$  takes the value 0.

$$x'_i = \begin{cases} 1 & \text{If } r < S(x_i), r \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Consequently, having a solution  $x_i$  encoded as a set of real numbers, the sigmoid function is used to transform the solution  $x_i$  into a set of probabilities that presents the chance for bit  $i$  to be flipping. The flipping chance is then used to compute the binary solution  $x'_i$ . Algorithm 2 presents the SBABO algorithm and Algorithm 3 presents the binarization algorithm

#### **Algorithm 2: SBABO Algorithm**

1. Objective function  $f(x) = (x_1, x_2, \dots, x_n)^T$
2. Initialization: randomly place buffaloes to nodes at the solution space;
3. Update the buffalo's fitness values using (1);
4. **Get the binary buffaloes using the Binarization Algorithm;**
5. Update the location of buffalo  $k$  ( $bpmax_k$  and  $bgmax$ ) using (2);
6. Is  $bgmax$  updating. Yes, go to 7. No, go to 2;
7. If the stopping criteria is not met, go back to algorithm step 3, else go to step 8.
8. Output the best solution.

#### **Algorithm 3: Binarization Algorithm**

**Input:** Real solution presentation  $x_i$

**Output:** Binary solution presentation  $x'_i$

For ( $i = 1$  to (problem size)) {

Calculate  $S(x_i)$  using (3);

If (random number  $r < S(x_i)$ )

$x'_i = 1$ ;

Otherwise

$x'_i = 0$ ;

}

### 3.2 LBABO ALGORITHM

In the LBABO algorithm, we propose to start the search by a binary population (the solutions are binary from the beginning) and replace the arithmetic operators used in the solution update equations (i.e. 1 and 2) by logical operators as follow:

1)  $val_2 - val_1$

Assuming two values  $val_1$  and  $val_2$ :

$$val_2 - val_1 = \begin{cases} val_2 & \text{if } val_2 \neq val_1 \\ \text{not}(val_1) & \text{otherwise (i.e. } val_2 = val_1) \end{cases}$$

2)  $val_2 + val_1$

$$val_2 + val_1 = \begin{cases} \text{not}(val_1) & \text{if } val_2 \neq val_1 \\ \text{and}(val_2, val_1) & \text{otherwise (i.e. } val_2 = val_1) \end{cases}$$

3) Coefficient \* (or /) val

$$\text{Coefficient} * (/) val = \begin{cases} val & \text{if coefficient} > r, r \in [0,1] \\ \text{not}(val) & \text{otherwise} \end{cases}$$

Coefficient can be  $lp_1$  or  $lp_2$ . Algorithm 4 presents the LBABO algorithm

#### **Algorithm 4. LBABO Algorithm**

1. Objective function  $f(x) = (x_1, x_2, \dots, x_n)^T$ ;
2. Initialization: randomly place buffaloes to nodes at the solution space **using binary values**;
3. Update the buffalo's fitness values using (1) **and logical operators**;
4. Update the location of buffalo  $k$  ( $bp_{max,k}$  and  $bg_{max}$ ) using (2) **and logical operators**;
5. Is  $bg_{max}$  updating. Yes, go to 6. No, go to 2;
6. If the stopping criteria is not met, go back to algorithm step 3, else go to step 7.
7. Output the best solution.

## 4. EXPERIMENTAL RESULTS

In order to investigate the performance of the proposed algorithms to solve hard binary optimization problems, we used some knapsack problem benchmarks of two knapsack problem versions: Single knapsack problem (KP) and multidimensional knapsack problem (MKP).

### 4.1 KP AND MKP PROBLEMS

The KP is a NP-hard problem [12]. Numerous practical applications of the KP can be found in many areas involving resource distribution, investment decision making, budget controlling, project selection and so one. The KP can be defined as follows: Assuming that we have a

knapsack with maximum capacity  $C$  and a set of  $N$  objects. Each object  $i$  has a profit  $p_i$  and a weight  $w_i$ . The problem consists to select a subset of objects that maximize the knapsack profit without exceeding the maximum capacity of the knapsack. The problem can be formulated as [12]:

$$\text{Maximize } \sum_{i=1}^N p_i x_i \quad (5)$$

$$\text{Subject } \sum_{i=1}^N w_i x_i \leq C \quad (6)$$

$$x_i \in \{0,1\}$$

Many variants of the KP were proposed in the literature including the MKP. MKP is an important issue in the class of KP. It is a NP-hard problem [13]. In the MKP, each item  $x_i$  has a profit  $p_i$  like in the simple KP. However, instead of having a single knapsack to fill, we have a number  $M$  of knapsack of capacity  $C_j$  ( $j = 1, \dots, M$ ). Each  $x_i$  has a weight  $w_{ij}$  that depends of the knapsack  $j$  (example: an object can have a weight 3 in knapsack 1, 5 in knapsack 2, etc.). A selected object must be in all knapsacks. The objective in MKP is to find a subset of objects that maximize the total profit without exceeding the capacity of all dimensions of the knapsack. MKP can be stated as follows [14]:

$$\text{Maximize } \sum_{i=1}^N p_i x_i \quad (7)$$

$$\text{Subject } \sum_{i=1}^N w_{ij} x_i \leq C_j \quad (8)$$

$$j=1, \dots, M \text{ and } x_i \in \{0,1\}$$

The MKP can be used to formulate many industrial problems such as the capital budgeting problem, allocating processors and databases in a distributed computer system, cutting stock, project selection and cargo loading problems [15]. Clearly, there are  $2^N$  potential solutions for these problems. It is obviously that KP and its variants are combinatorial optimization problems. Several techniques have been proposed to deal with KPs [12]. However, it appears to be impossible to obtain exact solutions in polynomial time. The main reason is that the required computation grows exponentially with the size of the problem. Therefore, it is often desirable to find near optimal solutions to these problems.

## 4.2 EXPERIMENTAL DATA

The proposed SBABO and LBABO algorithms were implemented in MATLAB R2014a. Using a laptop computer running Windows 7, Intel(R) Core(TM) i3-3110M CPU@ 2.40 GHz, 2.40GHz, 4GB RAM. The used parameters in the experiments are:

- Population size: 40.
- Iterations: 300.
- $lp1=0.7$ .
- $lp2=0.5$ .



Several experiments were performed to assess the efficiency and performance of our algorithms. In the first experiment, we have tested and compared our algorithms with a harmony search algorithm (NGHS) on some small KP instances taken from [16]. In the second experiment, we have used some big KP instances used in [8] to test and compare the proposed SBABO algorithm with the Binary Particle Swarm Optimization algorithm (BPSO) [17] which has a common point with the proposed SBABO algorithm. In fact, the two algorithms (SBABO and BPSO) used the Sigmoid Function to generate the binary solution. The used instances are six different instances with different problem sizes, in which the weights and profits are selected randomly. The different problem sizes  $N$  are 120, 200, 500, 700, 900 and 1000. In these instances, the knapsack capacity is calculated by using equation 9 [8]. The factor  $3/4$  indicates that about 75% of items are in the optimal solution.

$$C = \frac{3}{4} \sum_{i=0}^N w_i \quad (9)$$

In the third experiment, we have evaluated the performance of our algorithms on some MKP benchmarks taken from OR-Library. We have tested the proposed algorithms on some MKP instances taken from seven benchmarks named mknapi. The obtained results are compared with the exact solution (best known).

Finally, statistical tests of Freidman are carried out to test the significance of the difference in the accuracy of each method in this experiment. And the performances of the proposed algorithms (SBABO and LBABO) are also compared in terms of execution CPU time with the two problems (KP and MKP).

### 4.3. RESULTS AND DISCUSSION

Table 1 shows the experimental results of our algorithms (SBABO and LBABO) and the harmony search algorithm (NGHS) on ten KP tests with different sizes. The first column, indicates the instance name, the second column indicates the problem size, i.e. number of objects. The third and fourth columns indicate the obtained results by the SBABO and LBABO algorithms and the last column indicates the obtained results by the NGHS algorithm. Observation of the presented results in Table 1 indicates that the proposed discrete binary algorithms (i.e: SBABO and LBABO) perform well than NGHS algorithm in F6 test. The SBABO perform well than LBABO and NGHS algorithms in F8 test. And the three algorithms have the same results in the other instances.

Table 1.Experimental results with small kp instances

<i>Test</i>	<i>Size</i>	<i>SBABO</i>	<i>LBABO</i>	<i>NGHS</i>
F1	10	295	295	295
F2	20	1024	1024	1024
F3	4	35	35	35
F4	4	23	23	23
F5	15	481.0694	481.0694	481.0694
F6	10	52	52	50
F7	7	107	107	107
F8	23	9767	9767	9767
F9	5	130	130	130
F10	20	1025	1025	1025

Table 2 shows the experimental results of SBABO, LBABO and BPSO algorithms on big KP instances. The first column presents the problem size (i.e., instance). The second, third and fourth columns present the obtained results by the BPSO, SBABO and LBABO algorithms respectively. With each instance, the first line presents the best solutions and the second one presents the solution averages. The presented results show that the SBABO and LBABO algorithms outperform the BPSO algorithm, and the LBABO algorithm outperforms the SBABO algorithm. The statistical Friedman test in Figure 1 presents a comparison of the BPSO, SBABO and LBABO results. The LBABO algorithm ranks first in the Friedman test. The SBABO ranks second and BPSO ranks third. This statistical test shows that there is a significant difference between LBABO and BPSO algorithms.

Table 2. Experimental results with big kp instances.

Instance	BPSO	SBABO	LBABO
<b>120</b>	4296	4316	4504
	3840.8	4088.09	4357
<b>200</b>	7456	6778	7530
	5703	6480.56	7284.22
<b>500</b>	13116	14730	16853
	12471.2	14396.11	16174.25
<b>700</b>	18276	20501	23278
	17097.4	19348.07	22530.4
<b>900</b>	22857	24767	30196
	21736.6	24270.83	28864.5
<b>1000</b>	24933	27306	32948
	24050	26607.3	31936.86

Whereas, the difference between LBABO and SBABO results is not statistically significant. Consequently, the obtained results confirm that the proposed algorithms outperform the BPSO algorithm and prove that the proposed algorithms give good results.

Table 3 shows experimental results of the proposed algorithms over 7 instances of MKP problem. The first column indicates the instance index. The second and third column indicates the number of object and knapsack dimension, respectively. The fourth, fifth and sixth columns indicate the best known, the SBABO and the LBABO solutions, respectively. As we can see, the SBABO algorithm is able to find the best solution of the six first instances from the 7 instances. The LBABO algorithm is able to find the best solution of the five first instances from the 7 instances. The SBABO algorithm outperforms the LBABO algorithm on the two last instances (6 and 7). The statistical Friedman test in Figure 2 shows a comparison of the best known, SBABO and LBABO results. The SBABO ranks second after best known and LBABO ranks third. This statistical test shows that the difference between best known, LBABO and SBABO results is not statistically significant. Consequently, the obtained results confirm and prove that the proposed algorithms give good and promising results that can be considerably increased by the introduction of some specified knapsack heuristic operators using problem specific knowledge.

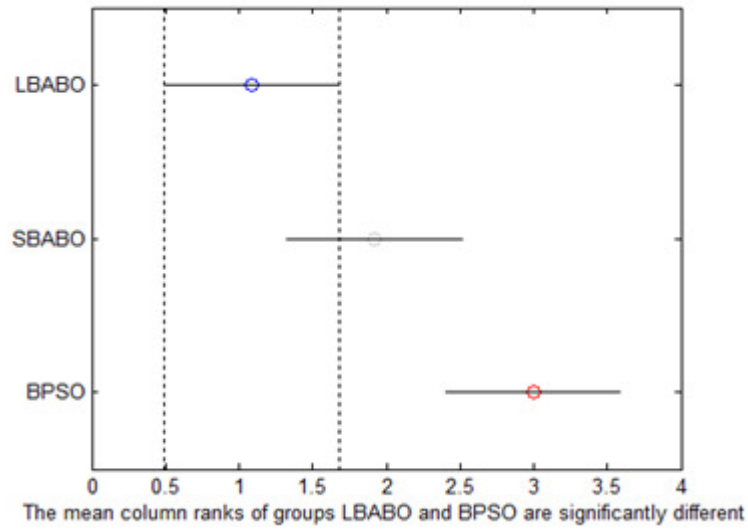


Figure 1. Friedman test compares SBABO, LBABO and BPSO on big KP instances.

Table 3. Experimental results of MKP with mknap1 instances

N°	N	M	Best Known	SBABO	LBABO
1	6	10	3800	3800	3800
2	10	10	8706,1	8706,1	8706,1
3	15	10	4015	4015	4015
4	20	10	6120	6120	6120
5	28	10	12400	12400	12400
6	39	5	10618	10618	10554
7	50	5	16537	16442	16371

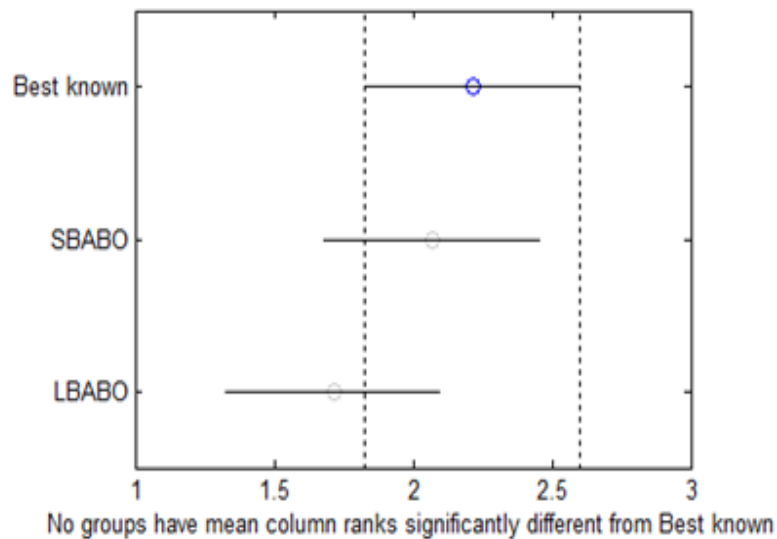


Figure 2. Friedman test compares SBABO, LBABO and best known on MKP instances.

Table 4 and 5 show a comparison of average computation time with KP and MKP instances, estimated by seconds, using a population of 40 solutions, 300 iterations with 5 executions of the programs. The obtained results are schematized in Figures 3 and 4. In terms of computing time, the obtained results do not show a big difference in execution time. In fact, in some instances SBABO is faster and in others LBABO is faster. In general, the two algorithms converge in the same interval time. This comes back to the fact that the two algorithms have the same body, only the phase of binarization of the solution that differs.

Table 4. Comparative CPU time with KP instances.

<b>Test</b>	<b>Size</b>	<b>SBABO</b>	<b>LBABO</b>
<b>F1</b>	10	2.50	2.75
<b>F2</b>	20	2.16	2.49
<b>F3</b>	4	1.67	1.48
<b>F4</b>	4	1.85	1.63
<b>F5</b>	15	2.03	2.73
<b>F6</b>	10	1.59	1.94
<b>F7</b>	7	2.08	2.30
<b>F8</b>	23	2.01	3.54
<b>F9</b>	5	2.02	2.14
<b>F10</b>	20	2.21	2.96

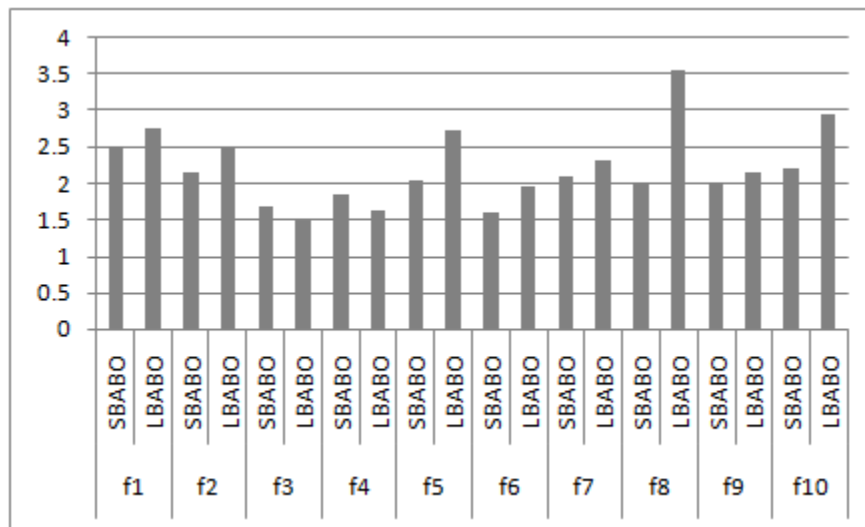


Figure 3. CPU time with KP instances

Table 5.Comparative CPU time with MKP instances

N°	n	M	SBABO	LBABO
1	6	10	3.09	2.99
2	10	10	3.38	2.72
3	15	10	3.49	3.30
4	20	10	3.55	4.28
5	28	10	3.75	5.11
6	39	5	4.97	5.45
7	50	5	5.95	5.83

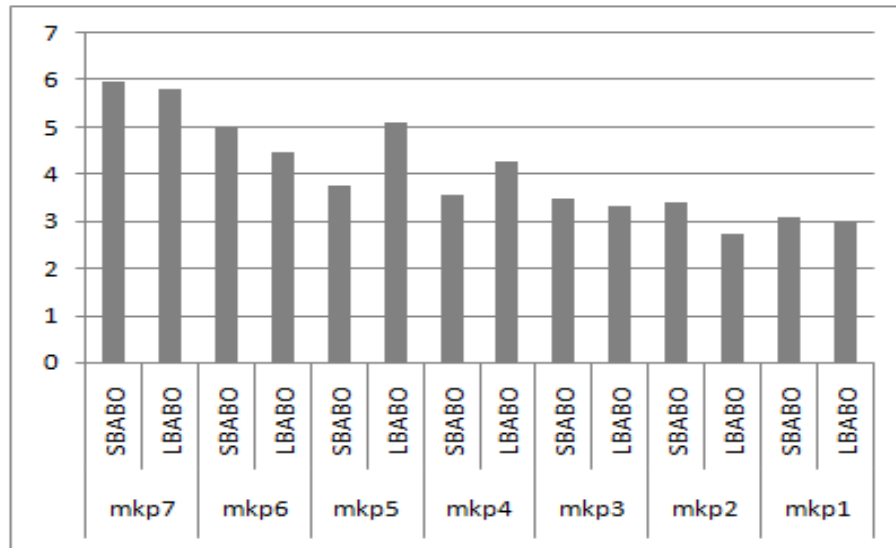


Figure 4.CPU time with MKP instances.

The ABO algorithm is a new swarm optimization algorithm. Considering its young age, there are few applications in optimization problems based on ABO algorithm. The main purpose of this paper is to validate that the ABO method is also effective for binary combinatorial optimization problems. That is why we proposed two discrete binary versions of ABO algorithm (called SBABO and LBABO) which led to two efficient ABO algorithm versions to deal with binary optimization problems.

During the different experiments, we noticed that SBABO algorithm explored the search space better than LBABO (see Figure 5). This comes down to the use of the Sigmoid Function and probability model to generate binary solutions. As shown in Figure 5, LBABO converges faster than SBABO which explains its results with MKP instances. It is notable that the performance of the algorithm is insensitive to their parameters such as  $lp_1$  and  $lp_2$ . These two parameters influence the good balance between exploration and exploitation of the search space. The diversity of the proposed algorithms is assured by the use of the elitism selection which guarantees that the best solutions are kept in each generation. The proposed algorithms can be implemented easily for other binary optimization problems.

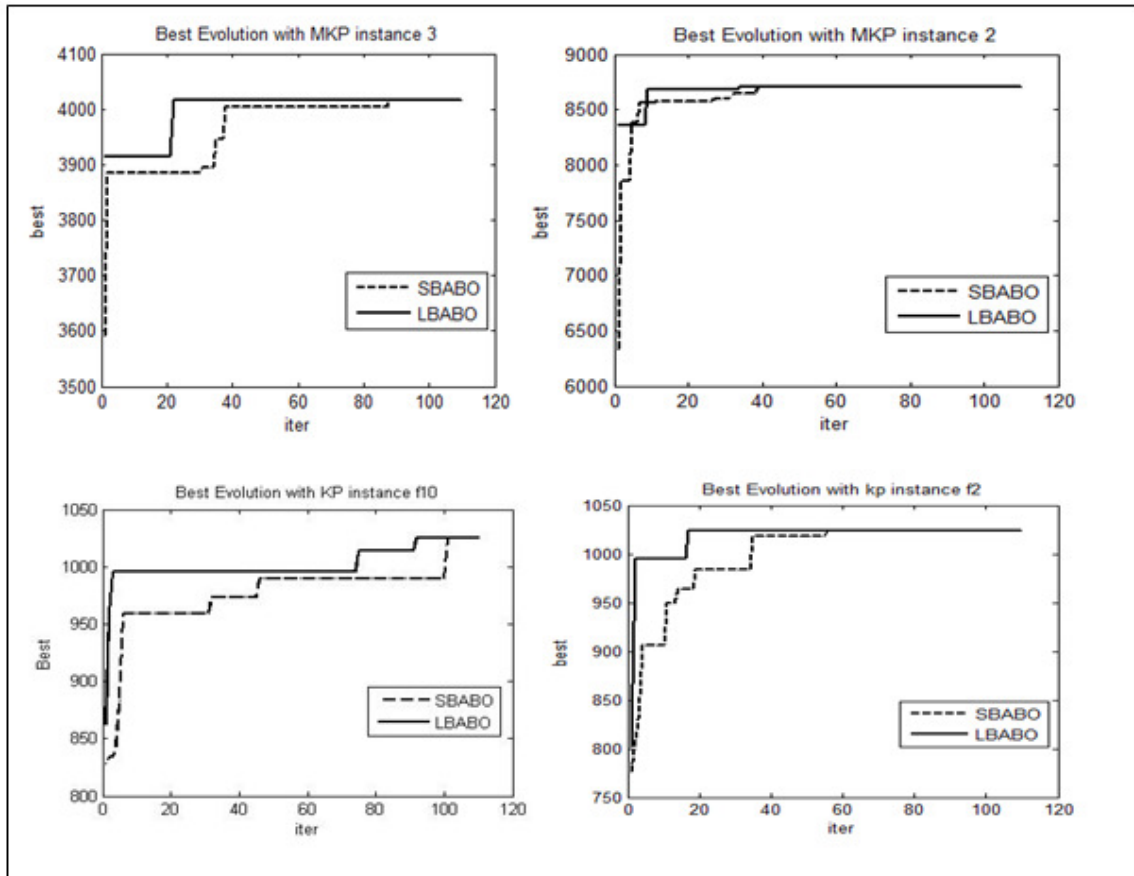


Figure 5. Evolution of best solution with KP and MKP using SBABO and LBABO

## 5. CONCLUSION AND PERSPECTIVES

In this paper, two discrete binary versions of African Buffalo Optimization algorithm are proposed. This contribution has two-fold aims: the first aim is to propose a binary version of ABO algorithm to deal with binary optimization problems. The second aim is to prove the effectiveness of the ABO algorithm in solving NP-hard combinatorial optimization problems. In the first version called SBABO we used the sigmoid function and probability model to generate binary solutions. In the second version called LBABO we used some logical operator to operate the binary solution. The proposed algorithms are used to solving two NP-hard binary combinatorial optimization problems: KP and MKP problems. The obtained results are compared with the harmony search algorithm (NGHS), the best known solution and the Binary Particle Swarm Optimization algorithm (BPSO) which has a common point with the proposed SBABO algorithm (the two algorithms used the sigmoid function). The experimental studies prove the feasibility and the effectiveness of the proposed algorithms. They proved that the proposed algorithms give good and promising results. However, there are several issues to improve the proposed algorithms. Firstly, in order to improve the performance of the proposed algorithms, we recommend integrating of a local search method in the algorithms core. In addition, hybridization with other operations inspired by other popular algorithms such as Genetic algorithm, Particle Swarm Optimization or Cuckoo Search will also be potentially fruitful. The proposed algorithms

can be also applied to solve many other binary optimization problems and real industrial problems.

## REFERENCES

- [1] Odili J.B, Kahar M.N.M, Anwar S. African Buffalo Optimization: A Swarm-Intelligence Technique. *Procedia Computer Science* 76. Elsevier, 2015. 443 – 448.
- [2] Odili J.B, Kahar M.N.M. Solving the Traveling Salesman's Problem Using the African Buffalo Optimization. *Computational Intelligence and Neuroscience*. Volume 2016, Article ID 1510256. Hindawi Publishing Corporation, 2015.
- [3] Odili J, Kahar M. N. M, Noraziah A and Kamarulzaman. S. F. A comparative evaluation of swarm intelligence techniques for solving combinatorial optimization problems. *International Journal of Advanced Robotic Systems*. DOI: 10.1177/1729881417705969. 2017.
- [4] Padmapriya R. Maheswari D. Channel Allocation Optimization using African Buffalo Optimization-Super Vector Machine for Networks. *Asian Journal of Information Technology*, 2017. DOI: 10.3923/ajit.2017.783.788.16: 783-788.
- [5] Liao C.J, Tseng C.T. and Luarn P. A discrete version of particle swarm optimization for flowshop scheduling problems. *Computers & Operations Research*. Elsevier, 2007. Vol. 34, No. 10, pp.3099–3111.
- [6] Huang S. H, Tian N, Wang.Y and Ji Z. Multi-objective flexible job-shop scheduling problem using modified discrete particle swarm optimization. *Springer Plus*, 2016. 5:1432.
- [7] Ammi M, Chikhi S. Cooperative Parallel Metaheuristics based Penguin Optimization Search for Solving the Vehicle Routing Problem. *International Journal of Applied Metaheuristic Computing*, 2016. Vol 7. Issue 1.
- [8] Gherboudj, A. and Chikhi, S. BPSO algorithms for knapsack problem. In Özcan, A., Zizka, J. and Nagamalai, D. (Eds.): *WiMo/CoNeCo, CCIS*, Springer. 2011. Vol. 162, pp.217–227.
- [9] Kong M. and Tian P. Apply the particle swarm optimization to the multidimensional knapsack problem. in Rutkowski, L. et al. (Eds.): *Proc. ICAISC 2006, LNAI*, Springer. 2006. Vol. 4029, pp.1140–1149.
- [10] Julstrom B.A. Greedy. Genetic, and greedy genetic algorithms for the quadratic knapsack problem. In *Proc. GECCO '05 Proceedings of the 2005 Conference on Genetic and Evolutionary Computation*, Publisher, ACM, 2005. pp.607–614.
- [11] Singh A. and Baghel A.S. A new grouping genetic algorithm for the quadratic multiple knapsack problem. In Cotta, C. and van Hemert, J. (Eds.): *Proc. EvoCOP 2007, LNCS*, Springer. 2007. Vol. 4446, pp.210–218.
- [12] Pisinger D. Where are the hard knapsack problems? *Computers and Operations Research*, 2005. Vol. 32, No. 9, pp.2271–2284.
- [13] Chu P.C, Beasley J.E. A genetic algorithm for the multidimensional knapsack problem. *Journal of Heuristics*, 1998. Vol. 4, No. 1, pp.63–86.

- [14] Angelelli E, Mansini R. and Speranza M.G. Kernel search: a general heuristic for the multi-dimensional knapsack problem. *Computers & Operations Research*, Elsevier, 2010. Vol. 37, No. 13, pp.2017–2026.
- [15] Vasquez M, Vimont Y. Improved results on the 0-1 multidimensional knapsack problem. *European Journal of Operational Research*, 2005. Vol. 165, No. 1, pp.70–81.
- [16] Zou D, Gao L, Li S. and Wu J. Solving 0-1 knapsack problem by a novel global harmony search algorithm. *Applied Soft Computing, The Impact of Soft Computing for the Progress of Artificial Intelligence*, March, 2011. Vol. 11, No. 2, pp.1556–1564.
- [17] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm. In *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics*, Piscataway, NJ, 1997. pp.4104–4109.

## **AUTHOR**

**Amira Gherboudj** is Senior Lecturer at the Algerian University of “Frères Mentouri, Constantine 1”. Dr. Gherboudj received her PhD degree in computer science in 2013 from the University of “Abdelhamid Mehri, Constantine 2”. Her research interests include combinatorial optimization methods and their applications to solve several problems from different domains.



# INTELLIGENT ELECTRONIC ASSESSMENT FOR SUBJECTIVE EXAMS

Alla Defallah Alrehily, Muazzam Ahmed Siddiqui, Seyed M Buhari

Faculty of computing and information technology, King Abdulaziz University,  
Saudi Arabia, Jeddah

## **ABSTRACT**

*In education, the use of electronic (E) examination systems is not a novel idea, as E-examination systems have been used to conduct objective assessments for the last few years. This research deals with randomly designed E-examinations and proposes an E-assessment system that can be used for subjective questions. This system assesses answers to subjective questions by finding a matching ratio for the keywords in instructor and student answers. The matching ratio is achieved based on semantic and document similarity. The assessment system is composed of four modules: preprocessing, keyword expansion, matching, and grading. A survey and case study were used in the research design to validate the proposed system. The examination assessment system will help instructors to save time, costs, and resources, while increasing efficiency and improving the productivity of exam setting and assessments.*

## **KEYWORDS**

*Subjective Assessments, E-Examination, WordNet, semantic similarity.*

## **1. INTRODUCTION**

With the rapid growth of modern education, the idea of E-learning system has been implemented to enhance the teaching of online courses, allowing instructors to offer online examinations through virtual classrooms. Electronic-learning overcomes many problems faced by students, such as the expense of traditional academic courses. Exams are an essential activity for students' learning as they assess the students' knowledge and level of understanding about a given subject. Therefore, the key aspects of an examination system are preparing a new paper for each student and conducting follow-up assessments.

In universities, a faculty member needs to set a minimum of three assessments per semester for a course (i.e., mid-term I, mid-term II, and final examination). Each faculty member generally teaches three courses per semester. Examination paper setting, and assessment are time- and labor-intensive, requiring many resources and placing immense pressure on course instructors. So, E-examination systems are importance in universities and institutions because it presents them electronic exams as a function open to all students in various places. For example, universities such as MIT, Berkeley, and Stanford have prepared electronic exams for massive open online courses (MOOCs) [1]. E-examination systems have the ability to check and set exam papers electronically, setting grades and assessing answers efficiently and yielding results quickly. These systems utilize fewer resources and minimal effort on behalf of the users. In contrast, traditional examination systems require physical resources such as pens and paper, greater user efforts, and more time.

Existing electronic-examination systems only evaluate exams with objective questions. But recently, researchers have identified the need to assess subjective questions using this tool [2]. Therefore, universities are in search of improved examination setting and assessment methods aside from the currently used manual method [3]. Therefore, there is a need for automatic examination and assessment systems in this context.

To tailor the existing assessment process in which examinations are set manually, this research aimed to develop an electronic assessment system for subjective examinations to assist instructors with exam setting and the assessment process. A new design is proposed for an electronic-examination assessment algorithm, which is achieved using the concept of semantic and document similarity to find a matching ratio between instructor and student answers to each question. The electronic system randomly generates exam papers, including both objective and subjective questions. A survey and case study are used in the research design to validate the electronic-examination system. In the case study, 10 students in King Abdul-Aziz University (KAU) were tested.

The rest of this paper is structured as follows. Section 2 discusses related research in the literature. Section 3 presents the problem statement. Section 4 explains the proposed system. Sections 5 and 6 describe the exam paper and the proposed assessment algorithm, respectively. In Section 7, the output of the examination assessment is presented, and in Section 8, the system is evaluated. Section 9 provides concluding remarks.

## 2. LITERATURE REVIEW

Xinming and Huosong [4] present an automated system that addresses the following problems with assessing subjective questions: synonymy, polysemy, and trickiness. Latent semantic analysis (LSA) and the ontology of a subject are introduced to solve the problems of synonymy and polysemy. A reference unit vector is introduced to reduce the problem of trickiness. The system consists of two databases: a science knowledge library and a question- and reference-answer library. The science knowledge library stores the ontology of a subject as text documents. The question- and reference-answer library stores questions as text documents and reference answers as a text document matrix. When a teacher adds new questions, a system using this science knowledge library will search for related points of knowledge and keywords and give them to the teacher. Then, the teacher will submit the reference answer to the system. It will process the reference answer using Chinese automatic segmentation, which produces text-document vectors and sends them to the teacher. Then, the teacher detects the terms and their weights for each vector and sends them back to the system. Weights of the terms in the reference answer are computed using the term-frequency and inverse-document-frequency functions. In the questions and reference answers, the library will save the vector of the reference answers and questions as text documents. To compute the similarity between a student's answer and the reference answer, the former is sent to the system, which assesses the answer using Chinese automatic segmentation and produces a text vector projected into k-dimensional LSA space. This LSA is formed by a vector using the mathematical technique of singular value decomposition (SVD), which represents terms and documents that are correlated with each other. The system computes the cosine similarity of student and reference answer vectors projected into k-dimensional LSA space in the reference unit vector.

In [5], machine learning techniques with and without ontology are presented to evaluate subjective answers. The techniques without ontology include LSA [4], generalized latent semantic analysis (GLSA), bilingual evaluation understudy (BLEU), and maximum entropy (MAXENT). Using ontology to evaluate subjective answers, student answers to questions and concept details are fetched from the ontology based on the type of question. If short questions are

answered, only a few details are extracted from the ontology. And if longer questions answered, details extracted from ontology are more and the similarity score among concepts extracted. After the information extracted from ontology, be configured a Multi-Hash map that used for evaluating answers. This Multi-Hash map collected all the words symmetrical for the same concept. If the concepts have a track among each other, then the length of such the track is computed. The authors combined ontology with machine learning techniques. The input of all machine learning techniques is the model answer and students' answers, a multi-hash map of Ontology concepts and distance among concepts. The method of combine ontology with machine learning techniques is constructing Ontology concepts of the sentences in the model answer and using the machine learning technique, merging concepts with the Ontology map. Using same machine learning technique, finding a correlation between every concept and students' answer in the multi-Hash map. To compute the final score of the similarity between students' answers with the model answer, the distance among the main concept and current concept is multiplied by the whole number of concepts having a positive correlation with students' answers. Then, this estimate is divided by a whole number of concepts in multi-Hash Map to construct final score. The most technique merged with Ontology is the word-weight technique. In this technique, the words are extracted from ontology and then words in the model answer are associated with ontology concepts. Finally, the weight of every keyword is computed.

Using the machine learning techniques without ontology, they take keywords of the model answer and student answer as input. The output is a similarity measure in the range between 0 and 1 where a value of 0 indicates no similarity and 1 indicates the high similarity. Before applying the machine learning techniques, pre-processing of words is tokenization, stop word removal, synonym search and stemming performed for the input.

Maram et al. [6] introduces an Automatic evaluation of an essay (AEE) system which is written in Arabic. The system presents a hybrid approach which integrates the LSA [4] and rhetorical structure theory (RST) algorithm. LSA method supports the semantic analysis of the essay, and the RST to evaluate the writing method and the cohesion of the essay. The LSA method finds the similarity ratio among two texts even if they do not include similar words. The system processes input essay into two phases is a training phase and testing phase. The training phase is made up of three parts: calculating the average of words per essay, calculating the most ten visible words on a given topic and applying LSA algorithm. The testing phase passes through a number of processes: 1) calculating LSA distance. 2) calculating the number of a vernacular. 3) calculating a number of repeated sentences. 4) calculating the length of the essay. 5) calculating number of spelling mistakes. 6) applying RST algorithm. 7) checking cohesion of essay related to the topic. Then applying two phases, the system computes the final score based on the cosine distance of LSA between the input essays and the training essays. The system graded school children essays based on three criteria which are 40% of the total score for writing method, 50% for the cohesion of the essay and 10% for spelling and grammar mistakes.

Anirudh et al. [7] propose an automated evaluation system for descriptive English answers that contains multiple sentences. The system evaluates the student's answer with an answer-key for questions of professional courses. It depends on a group of algorithms for natural language processing which are Wu and Palmer, Longest Common Substring (LCS), LSA [4], Cosine Similarity and Pure PMI-IR. The algorithms analyze the student's answer with an answer-key for finding the similarity score between them. Then, similarity scores extracted from algorithms are merged using the logistic regression machine learning to produce a score that is recommended by instructor. Wu-Palmer technique compares the word in the student's answer with each word in answer-key. If both words are present in the English dictionary, Wu-Palmer technique computes a similarity score for both words. Otherwise, if both words are not present in the dictionary, then the comparison is done using edit distance. LCS used to compare both sentences of the student's answer and answer-key. Then, the similarity score of LCS combined with a similarity score of

Wu-Palmer technique using the similarity matrix method. LSA uses SVD [4] on the similarity matrix that formed of both sentences. SVD produces two vectors representing two sentences. The similarity between two sentences is computed using cosine similarity. Pure PMI-IR combines all similarity scores of word pairs among sentences in one value using the similarity matrix method. The multi-class Logistic Regressors technique combines results of all five techniques to produce a score for the answer.

Ishioka and Kameda [8] propose an automated Japanese essay scoring system named jess. The system uses to mark essays in Japan for the University Entrance Exams. It assesses the essay from three metrics: rhetoric, organization, and content. Rhetoric means a syntactic variety that measures the details which are the ease of reading, diversity of vocabulary, percentage of big words and percentage of passive sentences. Organization means presenting and relating ideas in the essay. For organization assessment, jess examines the logical structure of the document and attempts to determine the occurrence of definite conjunctive expressions. Content means relevant information such as the precise information provided, and the vocabulary employed to the topic. For content assessment, jess applies a technique named LSA [4] which be applied to examine if the contents of a written essay react well with the essay prompt. Jess uses learning models which are editorials and columns extracted from the Mainichi Daily News newspaper.

In [9] proposes an approach of evaluation of online descriptive type students' answers using Hyperspace Analog to Language (HAL) procedure and Self-Organizing Map (SOM) method. To evaluate students' answer, the student writes the answer and sent as input to HAL. HAL constructs a high dimensional semantic matrix from a collection of an n-word vocabulary. Method for construct matrix through motivation a window of length "1" by the corpus through one-word increment. HAL ignores sentence boundaries, punctuation and converts each word to numeric vectors expressing information on its meanings for words. Inside window computes the distance between two words is "d", then computes "(1-d+1)" which denotes the weight of an association among two words. This matrix presents words by the analysis of lexical co-occurrence. Every word represents in the row vector based on the co-occurrence data for the words preceding this word and every word represents in the column vector based on the co-occurrence data for words following it. The matrix converts into a singular value by using SVD function [4]. Vector produced by HAL enters as an input to the Self-Organizing Map (SOM). It clusters words based on finding Euclidean distances denote the document similarity. SOM is neural technique. SOM takes vectors and produces a document map. Then, neurons are nearby will include the similar document. The authors compared SOM results with other clustering methods like Farthest First, Expectation Maximization (EM), Fuzzy c-Means, k-Means and Hierarchical. They concluded that SOM awards better performance.

kumaran and Sankar[10] propose a technique of an automated system for assessing the short answers using ontology mapping. Three stages of assessing the short answers are RDF sentence builder, ontology construction, and ontology mapping. In the first stage, the system constructs the RDF sentence for every sentence in student answer and model answer after reading the model answer and student answer as input in plaintext form. The system parses each sentence and builds the grammatical relationships each sentence. It uses Stanford typed dependency parser to represent dependency relationships. In the second stage, the RDF sentences are as input to ontology constructor to construct an ontology for them. The authors use sequential and coordinate links to construct RDF graph for the RDF sentences. The sequential link means that object or predicate is mutual among two RDF sentences. The coordinate link means that subject is same of two RDF sentences. Each link in ontology has the weight in the range of 0 to 100 based on the level of significance of that sentence in the answer and the whole weight of all links will be 100. In the third stage, the output of the previous stage is the model answer ontology and student answer ontology that use them the ontology mapping to perform matching operation. Output for this stage is the mark for the student answer depend on the weight age and the similarity score.

The method of the ontology mapping is the first finding the matching between the edges of the model answer ontology and student answer ontology are using the Cartesian product. The second, finding the similarity between two vertices of two ontologies using wordnet based similarity measure.

Raheel and Christopher [11] propose a system that provides a novel approach for automated marking of short answer questions. To compute the grade for the student's answer, authors introduce the architecture for the system that is composed of three phases to address the student's answer. Three phases are 1) spell checking and correction that is implemented by an Open Source spell checker like JOrtho. 2) parsing the student's answer using the Stanford Parser. This statistical parser can be creating parses with high accuracy. The parser offers the following results which are the part of speech tagged text and design dependency grammatical relations among singular words. 3) The Third phase of the processing answer is a comparison between the tagged text with syntactical structures specified by authors in Question and Answer Language. This phase addressed by syntax analyzer. Also, architecture contains analyzer of grammatical relation that compares between the grammatical relations in student answer with the grammatical relations specified by the examiner. The last task in the comparison phase is passing the results summarized from the syntax analyzer and the grammatical relation analyzer to the marker that calculates the final grade of the answer.

The Automatic marking system for a student's answer examination of the short essay was introduced by Mohd et al. [12]. The system applied to sentences were written using the Malay language that requires technique to process it. The technique mentioned in [11] which is the syntactic annotation and the dependency group to represent the Grammatical Relations (GR) from Malay sentences. To process the sentences from the marking scheme and the students' answers, all entries to the Computational Linguistic System (CLS) for linguistic processing like tokenizing, recognizing, collocating and extracting the GRs. The system contains a database for a table of Malay words and their Part of Speech (POS) to assist the CLS. To compute the mark for the student's answer, compare the GR extracted from the students' answers with the GR for the marking scheme. In other words, comparison components of the sentences as follows: subject to the subject, verb to the verb, object to object and phrase. The authors did the test of the system to view how the system gives marks compared to the marks awarded by a human. They selected lecturers have experienced in marking the scheme from Malaysia to set the mark for each question. The test presents which the system can give similar marks as marks awarded by the lecturers.

A new automated assessment algorithm for assessing the Chinese subjective answers was proposed by Runhua et al. [13]. The algorithm called Automated Word and Sentence Scoring (AWSS) assesses the student answers for the level of word and sentence. From fundamental problems of the Chinese, Natural Language Processing is the word segmentation, but this problem solved by the Institute of Computing Technology, the Chinese Lexical Analysis System (ICTCLAS). It assesses the student's answer to the standard answer in two phases as follows: 1) compute similarities between two words depend on How-Net. In this phase, they check keywords weight and phenomena of the synonym. The authors' present results of How-Net is satisfied. To compute the similarity between student answer with the standard answer for the level of the sentence, the authors' divide sentence to a series of words. Then computing the best matching pair of every word in the sentence and computing the sentence similarity as functions mentioned in [13]. 2) compute the similarity of sentences depending on dependency structure among words of a sentence. This phase parses the sentence by the language technology platform (LTP) to find out the dependency structure of the sentence. The method of computing dependency structure is finding a valid pair which is a noun, verb or subjective linking to the head of the sentence. Then, computing the sentence similarity based on dependency structure as functions mentioned in [13].

Xia et al. [14] design automatic scoring algorithm for a subjective question. They use the idea of a one-way approach degree depending on the closeness theory of fuzzy mathematics. The authors are calculating the closeness of two fuzzy sets which are set "A" denoted by the standard answer string and set "B" denoted by the student answer string. A fuzzy set is an ordered collection from a single character that decomposed from a string. To compute a one-way approach degree between two fuzzy sets "A" and "B", "B" contain n characters and one-way approach degree denoted by  $\delta(B, A) = m/n$  whereas m denotes by the effective sum number of the set B in each element in the set A.  $\delta(B, A)$  introduce B close to A unidirectional closeness. The introductory algorithm provides the aim of the system.

Zhenming et al. [15] propose a novel web-based online objective examination system for computer science education. This system conducts the examination and auto-marking of objective questions and operating questions. The system transmits answers and questions into the bit stream after encoding to ensure security and intrusion. It is the password protected system and camera are used to monitor the activities of students. The auto-grading system can automatically grade the answers, that are collected from the examination system. The objective questions can be graded effectively via fuzzy matching. But operating questions is difficult to grade by simple Matching technologies. Thus, researchers propose a universalized grading system that is achieved on the foundation of a database for key knowledge. The system does the following: first, they elicit all likely knowledge points and store them in a triple form (key, value, location). Then they make the question file via labelling the question point directly on it. After that, the system will add the identical question key to the standard key library. The last process of the system is comparing the answer file with the standard key library.

Our study is similar from previous studies for using the concept of semantic similarity and document similarity to find the matching ratio between instructor answer with student answer.

### 3. PROBLEM STATEMENT

In the education field, universities are currently setting and assessing the examination papers manually. Therefore, it is in need of automatic examination and assessment systems. Due to the manual exam setting and assessment for university faculties, they are facing following the main problems:

- 1) It is a tedious process to set exam papers and quizzes in every semester.
- 2) It needs a lot of time, more effort on instructors and consumes more resources to set and assess the examination papers especially if a number of students in the class are greater than thirty.
- 3) The paper-based examinations are currently scanned to convert them electronically for the review of The Accreditation Board for Engineering and Technology (ABET). This requires extra time, cost and resources.

To cater three issues, this research aimed to develop an electronic objective and subjective examination an assessment system to address the problems of universities and it will be helpful for the other universities inside and outside of the Kingdom of Saudi Arabia. It is anticipated that the proposed system will help instructors in the exam setting and its assessment. The proposed system will save time, cost, resources, increase efficiency and improve the productivity of exam setting and assessments.

#### 4. THE SYSTEM USERS

The electronic system consists of two concepts which are the examinations setting and assessment system. Figure 1 below shows the electronic system users. The system has two courses which are System Analysis and Design and Software Engineering. The questions and answers of exams in two courses collected and questions composed of equally distributed simple, average and difficult questions.

The main users of a system which are an instructor, head of the track, head of department, student and system administrator. Each of users has own screen to log in with the user name and password. The users have specific functions applied to them in the system. The instructor can create the objective and subjective questions and select course type and write the grade for each question. He can approve grades for students on the main screen. The head of the track can modify and approve all the objective and subjective questions which are created by the instructor. The student selects a course to start the exam and solve the questions. A student can view final grade. After approving the instructor for final grades, head of the department can approve and publish it to students.

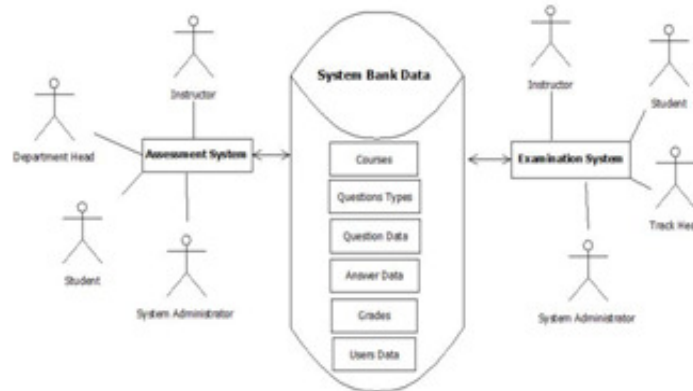


Figure 1. System users

#### 5. EXAM PAPER DESIGN

The Exam paper had consisted of two sections. The first section was objective questions of multiple choices. The second section was subjective/ descriptive questions which are a short essay, definitions, and lists. Questions were coming one after the other -in both parts. The system selected questions randomly of the database using function "RAND". In the system, two courses were System Analysis and Design, Software Engineering. The student selected the course to start the exam. Then, the system selected randomly five objective questions and five subjective questions. The system had a specific time for the exam. It set 25 minutes for solving objective section and 55 minutes for solving subjective section.

#### 6. THE PROPOSED ASSESSMENT ALGORITHM

The Assessment system architecture consists of different modules which assess student answers with reference answers. The modules are a pre-processing module, Keyword Expansion module, matching module and Grading Module. Figure 2 below presents subjective examinations assessment algorithm. The details were explained in next section.

### 6.1. Preprocessing module

The inputs to the module were the reference answers provided by the instructor and student answers. Two answers converted to lowercase using “lower ()”. Then, the module removes stop words, punctuations, and prepositions from converted answers. The output of the module is the cleaned answer.

For processing punctuations of text, the module was called a string containing all characters considered punctuation. This is the string `"'\"#$%&'()*+,-./:;<=>?@[\\]^_`{|}~"''''`. The module has converted the string of punctuations to set using “set ()”. If the characters of the answer were not in punctuations set, the module has joined the characters with an empty space using “join ()”. The sentence can be split into words using the method “word tokenize ()”. Tokenizers can be used to find the words in a string. The module imports natural language toolkit(NLTK) can provide “word tokenize ()” and other methods for processing texts. The output of the method is a list of words named keywords.

For processing stop words of text, the module was imported stop words of the NLTK package can provide a list of stop words. If keywords were not in a set of English stop words, the keywords remain in the list.

For processing prepositions of text, the module was classified keywords into their parts of speech is known as part-of-speech tagging (POS-tagging). Parts of speech are also known as lexical categories. The module was imported POS-tagging of the NLTK package can attach a part-of-speech tag to each word. It was set POS-tagging using the method “pos tag ()”. If POS-tagging of keywords were not in prepositions list, the keywords remain in the list. After applying all processing texts methods, the module returned all keywords of answer joined with an empty space using “join ()”.

As for the objective answers are being processed by fetching the reference answers from Objective Questions Approved store and also fetches the student answers from Exam Session store. If student answer matches with reference answer, then answer is correct. Otherwise, the answer is not correct.

*The inputs to the pre-processing module:*

Reference answer="The Cost required to deVelop the System"

Student answer = "The scheme NEED to involve by estimate the cost"

*The outputs of the pre-processing module:*

Cleaned Reference answer =" cost required develop system"

Cleaned Student answer =" scheme need involve estimate cost"

### 6.2. Keyword Expansion module

After cleaning student's answer and reference answer in the Pre-processing module. The cleaned reference answer split to a list of keywords using “split ()”. The Keyword Expansion module achieved two tasks were got synonyms and apply synonyms. The first task was got synonyms set each keyword of wordnet. The module was imported wordnet of the NLTK package can provide synonyms set of English words. Look up a keyword using function “synsets()”. The output of this process is synonyms set of keywords. It was accessed synonyms set and get the synonyms each keyword using function “lemma names ()”. The synonyms contained within synonyms set are called lemmas. The module was imported chain from iterator tools (itertools) module makes an iterator that returns the synonyms from the many iterators until it is exhausted. Thus, synonyms presented as the chain using the function “chain. From iterable()”. Then, the module was put



synonyms in the set using “set ()”. The output of the first task is synonyms of the keyword. Then, it was presented each keyword and its synonyms as a dictionary using the method “Dict ()”.

After finding synonyms of keywords of reference answer. The second task received two inputs were cleaned student's answer and dictionary of keywords in reference answer. The cleaned student's answer split to a list contained in words named text. The module was generated empty list named words list. Each word in the text added to words list using the method “append ()”. The module was called items of keywords and synonyms of reference answer using the method “items ()”. Thus, this method returned (keyword, synonym) tuple pairs. If the word in the student's answer matched with the synonym as presented in tuple pair. Then, this module has deleted this word from words list using delete operator. And it was replaced its place the basic keyword as presented in tuple pair using “append ()”. The basic keyword is from the reference answer. The output of the second task is the basic keywords and words have not the synonym. All these words joined with an empty space using “join ()”. Thus, this is the new student's answer after converting synonyms to the basic keywords. This task applied also to cleaned reference answer.

*The inputs to the Keyword Expansion module:*

Cleaned Reference answer=” cost required develop system”

Cleaned Student answer=” scheme need involve estimate cost”

*a. get synonyms of keywords of cleaned reference answer*

*input to task a:*

Cleaned Reference answer=” cost required develop system”

*Output of task a:*

```
'system': {'system', 'organization', 'arrangement', 'scheme', 'organisation', 'system_of_rules'}
, 'develop': {'germinate', 'grow', 'explicate', 'acquire', 'rise', 'get', 'modernize', 'break', 'evolve',
'make_grow', 'recrudesce', 'develop', 'originate', 'uprise', 'modernise', 'build_up', 'produce',
'educate', 'arise', 'prepare', 'train', 'formulate', 'spring_up'}, 'cost': {'price', 'be', 'monetary_value',
'cost', 'toll'}, 'required': {'requisite', 'mandatory', 'need', 'take', 'necessitate', 'require', 'involve', 'ask',
'command', 'needful', 'call_for', 'want', 'required', 'postulate', 'demand', 'expect', 'compulsory',
'needed'}
```

*b. applies synonyms of cleaned reference answer and cleaned student answer*

The inputs to task b:

1.Cleaned Reference answer=” cost required develop system”

2.Cleaned Student answer=” scheme need involve estimate cost”

3.dictionary of task a

The outputs of task b:

1. Cleaned Reference answer

“cost” =” cost”

“required” = “required”

“develop” = “develop”

“system” =” system”

2. Cleaned Student answer

“scheme” =” system”

“need” =” required”

“involve” = “required”

“estimate” did not match any synonym

“cost” =” cost”

*The outputs of the Keyword Expansion module:*

New Reference answer=” cost required develop system”

New Student answer=” system required required estimate cost”

### 6.3. Matching module

This module achieved two tasks: the first task was converted text to vector and the second task was computed cosine similarity between two vectors. The first task received two inputs were the new student's answer and new reference answer from the previous module. It displays the textual representation of two answers into Vector Space Model (VSM). VSM represents answers as vectors in n-dimensional space where n is the total number of keywords in all the answers.

The first task works as the following:

1. import the module regular expressions defined as re. This module provides an interface to the regular expression engine. It can compile the regular expression to pattern objects which have methods for pattern matches.
2. construct the regular expression as pattern object named WORD= re.compile(r'\w+').The regular expression r'\w+' passed to the object as a string. It used to match words in the target answer.
3. The module was applied directly the method “findall” on the regular expression object “WORD”. The method received the target answer to finding matches in.
4. The output of step 3 is a list contains on matched keywords named keywords.
5. To compute how many frequencies of matched keywords and not matched keywords each answer. The module was imported Counter from collections module and call the method “counter ()” which receives keywords list of step 4. It counts of keywords per the answer. The output of this process is keyword frequency vectors are created each student's answer and reference answer. In keyword frequency vector, the keyword was indicated key and frequency number was indicated value. The mathematical expression of step 5 represents as follows whereas Keyword (K), Keyword Frequency (KF), frequency (fr) and Answer Vector(AV):

$$KF(K,A) = \sum_{x \in A} fr(x,K) \text{ where } fr(x,K) = 1 \text{ if } x = K \text{ otherwise } fr(x,K) = 0$$

$$AV = (KF(K1,A), KF(K2,A), \dots, KF(Kn,A)) \text{ where } n \text{ is number of keywords} \quad (1)$$

The basic trend in the research is finding matched keywords between the student's answer vector and reference answer vector to assess the student's answer is correct or incorrect. This is the second task received two inputs were the student's answer vector and reference answer vector. It computed similarity ratio between two vectors using the similarity method is known document similarity. It used measure named cosine which computes the distance angle between the student's answer vector and reference answer vector. The mathematical expression of the task cosine similarity is as follows. Whereas it gives two vectors are Reference Answer Vector(RAV) and Student Answer Vector(SAV):

$$\text{similarity ratio} = \cos(\theta) = \frac{\sum_{i=1}^n RAV_i \times SAV_i}{\sqrt{\sum_{i=1}^n RAV_i^2 \times \sum_{i=1}^n SAV_i^2}} \quad (2)$$

The programmatic representation of the function “cosine similarity” is as the following:

1.To represent numerator as in the mathematical expression, extracting all the keywords of each vector using the method “keys” which returned keywords list of RAV and SAV. Then, putting keywords list of each vector as a set. Finding the intersection between two sets to extract matched keywords. The intersection represented as in the programmatic representation set (RAV. keys) & set (SAV. Keys). Each value of RAV in the intersection multiplied by each value of SAV. Then, summation all results using the method “sum”.

2. To represent denominator as in the mathematical expression, square each value in RAV then summation all values. And also, the module was applied to SAV values. The result for RAV named “sum 1” and for SAV named “sum 2”. After that, the module was extracted sqrt of “sum 1” and “sum 2”. The result of sqrt 1 multiplied by the result of sqrt 2.

3.If the result of numerator equals to zero, similarity ratio was zero. Otherwise, similarity ratio was the result of divide numerator to the denominator.

The similarity ratio of function “cosine similarity” represented the cosine of the angle was between 0 to 1. The similarity ratio was 1 means student answer is matched with reference answer. If the similarity ratio is the approximate number closer to 1 means that student answer is more similar for reference answer. Otherwise, if the approximate number closer to 0 means student answer is less similar for reference answer. Similarity Percentage (SP)is calculated as follows:

$$SP\% = \text{similarity ratio} \times 100 \quad (3)$$

*The inputs to the matching module:*

New Reference answer=” cost required develop system”

New Student answer=” system required required estimate cost”

*a. convert text to vector*

reference answer vector ({'system': 1, 'develop': 1, 'cost': 1, 'required': 1})

student answer vector ({'required': 2, 'cost': 1, 'system': 1, 'estimate': 1})

*b. compute cosine similarity*

cosine similarity ratio between RAV and SAV =0.75

*The outputs of the matching module:*

SP=75%

## 6.4. Grading Module

The module computes the full mark based on the similarity percentage. The full mark of the exam paper is ten out of ten. The number of exam questions is ten questions, and each question is of one mark. After processing answers using previous modules, this module gets the percentage of similarity of each question. If the similarity percentage between the student's answer and reference answer are getting between 70% and 100%, the student's answer is correct and full mark awarded. Otherwise, it is incorrect. Then the module computes the final grade of the exam by collecting grades of all questions. And grades store in Exam Session store and Grades Not Approved store.

Figure 2. subjective examinations assessment algorithm

**Input 1: student answer**

**Input 2: reference answer**

**The output of all modules: similarity ratio**

## 1. Pre-processing module

### *a. removes punctuations*

```
exclude = call all characters considered punctuation
for a character in two answers
    if character not in exclude
        sentence = join all characters with an empty space
    End if
tokens = split words in a sentence into tokens
End for
```

### *b. removes stop words*

```
Stop words = call list of English stop words
for word in tokens
    if word not in Stop words
        tokens = put a word in the list.
    End if
End for
```

### *c. removes prepositions*

```
tagged = call parts of speech tagging and define it to tokens
for a tag in tagged
    if tag not in prepositions list
        keywords = put tokens in the list
    End if
End for
Output 1: cleaned student answer
Output 2: cleaned reference answer
```

## 2. Keyword Expansion module

Input 1: cleaned student answer

Input 2: cleaned reference answer

### *a. get synonyms of keywords of cleaned reference answer*

```
synonyms set = call synonyms set of keywords
for a keyword in synonyms set
    synonyms = get synonyms each keyword as chain
    keywords and synonyms of reference answer = present each keyword and its synonyms as a
    dictionary
End for
```

### *b. applies synonyms*

```
text = split cleaned student answer to list
words list = generate an empty list
Tuple pairs (keyword, synonym) = call items of keywords and synonyms of reference answer
for word in the text
```

```

add a word to words list
for keyword and synonym in tuple pairs
    if the word in the synonym
        delete word of words list
        add a keyword to words list
    End if
End for
End for
Output 1: new student answer after applying synonyms
Output 2: new reference answer after applying synonyms

```

### 3. Matching module

Input 1: new student answer  
 Input 2: new reference answer

#### *a. convert text to vector*

```

words = find keywords in new student answer
SAV= count words
RAV= count words

```

#### *b. compute cosine similarity*

```

SAV keywords = call keys of SAV
SAV set = put SAV keywords in the set
RAV keywords = call keys of RAV
RAV set = put RAV keywords in the set
intersection = Find the intersection between two sets
for x in the intersection
    value 1 = get x value of SAV
    value 2 = get x value of RAV
    values = value 1 * value 2
    numerator = summation all values
End for
for x in SAV keywords
    square 1 = square x value of SAV
    sum 1 = summation all square 1
End for
for x in RAV keywords
    square 2 = square x value of RAV
    sum 2 = summation all square 2
End for
Sqrt 1 = sqrt of sum 1
Sqrt 2 = sqrt of sum 2
denominator = Sqrt 1 * Sqrt 2
if not denominator
    return 0
else
    similarity ratio = numerator / denominator
return similarity ratio
End if

```

## 7. OUTPUT OF EXAMINATION ASSESSMENT

The electronic system interfaces of users and exam paper implemented using PHP and java script languages. The system database implemented using my SQL. The assessment algorithm for exam paper implements using python program. After the student has solved exam, student answers with reference answers evaluated using assessment algorithm. Figure 3 shows bellow assessment results of the exam paper.



Figure 3. Assessment Output of the Exam Paper.

## 8. EVALUATION

Electronic system evaluated in several respects. First, in terms, the quality evaluation of the system will be using a survey. Second, in terms the performance evaluation using popular evaluation measures for electronic assessment and traditional assessment. And, evaluation using Spearman correlation for electronic assessment with traditional assessment.

### 8.1. The system quality Evaluation

The survey will be used research design to validate the quality of the electronic system. The survey distributed on different categories of KAU and other universities. The categories are the users of the system. The sample size is fifty user which tested the quality of the system. The fifty users are twenty-one students, six heads of departments, eight heads of tracks and three administrators of systems. Most users are female, and others are male. When analysing the survey results, 86% of users strongly agree and 14% of users agree which the system gives correct results when they used it. And also, same previous percentages, they take less time for accomplishing any task into the system. 78% of users strongly agree and 22% of users agree which the efficiency of the system is fast to accomplish their tasks. 98% of users strongly agree and 2% of users agree

which the system is easy to accomplish their tasks when they used for the first time. 96% of users strongly agree and 4% of users agree which functions as users of the system is completed. All users of the system are 100% strongly agree for consuming less cost and resources at use the system and accomplishing tasks. And they are satisfied with the whole system. Thus, the system achieves the main criteria of the system evaluation that it consumes a less of time and fewer resources. And, it reduces the effort on the users of the system and functions are high quality.

## 8.2. Comparison Electronic Assessment with Traditional Assessment Result

The electronic system has experimented on ten students of KAU. Data sets are 100 questions composed of 50 objective questions and 50 subjective questions of different courses. Experimental results of answers grades are conducted and analyzed using excel program. The study was used commonly the evaluation measures which are recall, precision, accuracy and F measure for measuring retrieval effectiveness of electronic assessment and traditional assessment. Traditional assessment is using instructors approach to assessing exam papers manually. The instructor's approach counts the correct words in the student answer and divides them by the total number of words in instructor answer. Then multiply the result at 100 to extract the similarity percentage between student answer with instructor answer manually.

The table following shows the mathematical expression of previously defined measures whereas True Positive (TP) is the number of answers correctly labeled as positives, True Negative (TN) is the number of answers correctly labeled as negatives, False Positive (FP) is the number of answers incorrectly labeled as positives and False Negative (FN) is the number of answers incorrectly labeled as negatives.

Table 1. Mathematical Expressions of Evaluation Measures

Measure	Mathematical Expression
Recall	$TP / (TP + FN)$
Precision	$TP / (TP + FP)$
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
F measure	$(2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision})$

Figure 4 shows the Performance comparison for electronic and traditional assessments using common evaluation measures. The recall is called True Positives Rate in the electronic assessment was 0.9778 and was 0.9655 in traditional assessment. Precision is called Positives Predictive Value was 0.9072 in electronic assessment and was 0.9882 in traditional assessment. Accuracy is called true results was 0.8900 in electronic assessment and was 0.9600 in traditional assessment. Finally, F measure is the geometric mean of recall and precision. It was 0.9412 in electronic assessment and was 0.9767 in traditional assessment.

The study was concluded recall rate for electronic assessment is so much closer to traditional assessment. When compared precision, accuracy and F measure for electronic assessment with traditional assessment, the study was concluded these measures for traditional assessment is higher than electronic assessment.

For evaluation purpose, the Spearman correlation is very important to note the extent of correlation between electronic assessments with traditional assessments. The Spearman correlation computed using excel program. The results of Spearman correlation were 0.83. Thus, the correlation represented the strength of the relationship between grades computed by electronic assessment and traditional assessment.

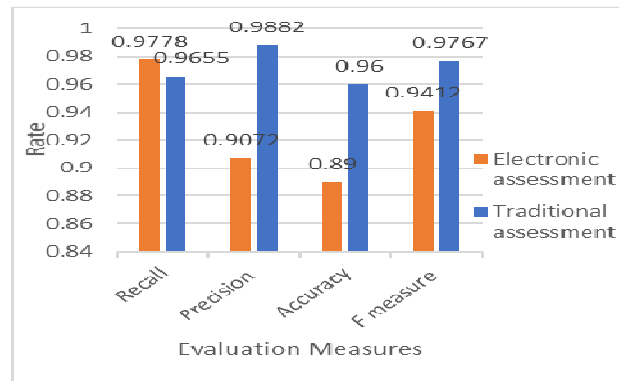


Figure 4. The Performance comparison for electronic and traditional assessments.

## 9. CONCLUSIONS

In the education domain, electronic examination systems are used to deal with objective assessments. Now, we need electronic examination systems to assess subjective questions in exams. There are several problems associated with the manual examination and assessment processes such as time-consuming, costly, enormous resources, a lot of efforts and huge pressure on instructors.

The paper has introduced a new design for an electronic examinations assessment system which achieves using the concept of semantic similarity and document similarity to find matching between instructor answer with student answer for each question. Then the system extracts the grade based on a percentage of similarity. The electronic grades correlate with instructor grades using Spearman's correlation. The accuracy of assessment using the electronic system is high. Thus, the proposed system will be beneficial for the faculties of other universities inside and outside of Kingdom of Saudi Arabia. The electronic system will help instructors in the exam setting and its assessment. It will save time, cost, resources, increase efficiency and improve the productivity of exam setting and assessments. Future work will develop assessment algorithm to address syntax errors of keywords and investigate high equality and performance for assessing them.

## REFERENCES

- [1] J. Dreier, R. Giustolisi, A. Kassem, P. Lafourcade, G. Lenzini, and P. Y. A. Ryan. Formal analysis of electronic exams. In *SECRYPT'14*. SciTePress, 2014.
- [2] P. Kudi, A. Manekar, K. Daware, and T. Dhattrak, "Online Examination with short text matching," in *Wireless Computing and Networking (GCWCN)*, 2014 IEEE Global Conference on, 2014, pp. 56–60.
- [3] K. Woodford and P. Bancroft, "Multiple choice questions not considered harmful," in *Proceedings of the 7th Australasian conference on Computing education-Volume 42*, 2005, pp. 109–116.
- [4] X. Hu, and H. Xia, "Automated Assessment System for Subjective Questions Based on LSI," *Third International Symposium on Intelligent Information Technology and Security Informatics*, Jinggangshan, China, pp. 250-254, April 2010.
- [5] M. S.Devi and H.Mittal, "Machine Learning Techniques With Ontology for Subjective Answer Evaluation," *International Journal on Natural Language Computing*, Vol. 5, No.2, April 2016.



- [6] M.F. Al-Jouie, A.M. Azmi, "Automated Evaluation of School Children Essays in Arabic," 3rd International Conference on Arabic Computational Linguistics, 2017, vol.117, pp.19-22.
- [7] A.Kashi, S.Shastri and A. R.Deshpande, "A Score Recommendation System Towards Automating Assessment In Professional Courses," 2016 IEEE Eighth International Conference on Technology for Education, 2016, pp.140-143.
- [8] T. Ishioka and M. Kameda, "Automated Japanese essay scoring system: Jess," in Proc. of the 15th Int'l Workshop on Database and Expert Systems Applications, 2004, pp. 4-8.
- [9] K.Meena and R.Lawrance, "Evaluation of the Descriptive type answers using Hyperspace Analog to Language and Self-organizing Map", Proc. IEEE International Conference on Computational Intelligence and Computing Research, 2014, pp.558-562.
- [10] Vkumaran and A Sankar, "Towards an automated system for short-answer assessment using ontology mapping," International Arab Journal of e-Technology, Vol. 4, No. 1, January 2015.
- [11] R. Siddiqi and C. J. Harrison, "A systematic approach to the automated marking of short-answer questions," Proceedings of the 12th IEEE International Multi topic Conference (IEEE INMIC 2008), Karachi, Pakistan, pp. 329-332, 2008.
- [12] M. J. A. Aziz, F. D. Ahmad, A. A. A. Ghani, and R. Mahmood, "Automated Marking System for Short Answer examination (AMSSAE)," in Industrial Electronics & Applications, 2009. ISIEA 2009. IEEE Symposium on, 2009, pp. 47-51.
- [13] R. Li, Y.Zhu and Z.Wu, "A new algorithm to the automated assessment of the Chinese subjective answer," IEEE International Conference on Information Technology and Applications, pp.228 – 231, Chengdu, China, 16-17 Nov. 2013.
- [14] X.Yaowen, L.Zhiping, L.Saidong and T.Guohua, "The Design and Implementation of Subjective Questions Automatic Scoring Algorithm in Intelligent Tutoring System," 2nd International Symposium on Computer, Communication, Control and Automation, Vols. 347-350, pp. 2647-2650, 2013.
- [15] Y. Zhenming, Z. Liang, and Z. Guohua, "A novel web-based online examination system for computer science education," in 2013 IEEE Frontiers in Education Conference (FIE), 2003, vol. 3, pp. S3F7–10.

## AUTHORS

**Alla Alrehily** is a student in King Abdul-Aziz University, Saudi Arabia. She teaches master of Faculty of Computing and Information Technology. Her research interests intelligent information retrieval.

**Muazzam Siddiqui** is an Associate Professor at the Faculty of Computing and Information Technology, King Abdul-Aziz University. He received his BE in Electrical Engineering from NED University of Engineering and Technology, Pakistan, and MS in Computer Science and PhD in Modelling and Simulation from the University of Central Florida, USA. His research interests include sentiment analysis, named entity recognition, and keyword and relationship extraction.



**Sayed Buhari** is an Associate Professor in Faculty of Computing and Information Technology, King Abdul-Aziz University, Saudi Arabia. His current research interests are in the areas of cognitive radio networks, grid computing, IPv6 performance testing and high-performance computing.



*INTENTIONAL BLANK*

# DYNAMIC PHONE WARPING – A METHOD TO MEASURE THE DISTANCE BETWEEN PRONUNCIATIONS

Akella Amarendra Babu<sup>1</sup>, and Ramadevi Yellasiri<sup>2</sup>

<sup>1</sup>St. Martin's Engineering College, Dhulapally, Secunderabad, India

<sup>2</sup>CBIT, Hyderabad, Telangana, India

## ABSTRACT

*Human beings generate different speech waveforms while speaking the same word at different times. Also, different human beings have different accents and generate significantly varying speech waveforms for the same word. There is a need to measure the distances between various words which facilitate preparation of pronunciation dictionaries. A new algorithm called Dynamic Phone Warping (DPW) is presented in this paper. It uses dynamic programming technique for global alignment and shortest distance measurements. The DPW algorithm can be used to enhance the pronunciation dictionaries of the well-known languages like English or to build pronunciation dictionaries to the less known sparse languages. The precision measurement experiments show 88.9% accuracy.*

## KEYWORDS

*Natural Language processing, word distance measurements, pronunciation dictionaries.*

## 1. INTRODUCTION

Pronunciation dictionaries are not available for all languages and the accents of various regions. This paper aims to build online pronunciation dictionaries using sound distance measurements. Human beings hear a word; compare it with the words in the memory and select the word which highest similarity to the input word. The objective of this paper is to follow the technique adopted by the human beings and prepare the pronunciation dictionaries. The primary focus of this paper is to measure distances between and sounds and to use this data to measure the distances between the words.

The reasons for the pronunciation variability are as under:

**1.1 Speaker's Accent:** The accent of the speaker depends on his mother tongue [1, 2]. The difference is negligible in respect of the speakers of the same country. But the difference is glaring in respect of foreign speakers.

**1.2 Speaker's Emotions:** The pronunciation of the same word would be different when spoken with different emotions like joy, love, anger, sadness and shame [3, 4].

**1.3 Speaking Style:** The speaker style varies when speaking to various people. The same name is spoken with different pronunciation while addressing an office peon and while addressing your friend.

**1.4 Speech Disfluencies:** There will be lot of gaps and filler sounds while speaking. It interrupts the normal of the human beings. This phenomenon creates pronunciation variability [5].

The natural speech results in generating different formant frequencies for the same spoken phoneme due to above reasons. Therefore, the phoneme sequences generated for a word will vary and depend on the speaker's accent, mood and the context [6].

The next section reviews the literature related to this work. Section three covers the theoretical background to the proposed algorithm of Dynamic Phone Warping (DPW). Section four covers the measurement of distance between various phonemic sounds produced by human beings. DPW algorithm is described in section five. Experimental details and analysis of results are discussed in section six. Some of the applications which can be developed based on the DPW methods of phoneme distance measurements are discussed in section seven.

## 2. RELATED WORK

The methods proposed for preparation of pronunciation dictionaries are discussed in this section. Pronunciation dictionaries are manually generated using linguistic knowledge are covered knowledge based methods. They are Grapheme-to-Phoneme (G2P) and Phoneme-to-Phoneme (P2P) conversions.

Stefan Hahn, Paul Vozila and Maximilian Bisani have used G2P methods for comparing large pronunciation dictionaries [7]. Algorithms are developed for grapheme to phoneme translation in [8]. It is used in applications used for searching the databases and speech synthesis. M. Adda-Decker and L. Lamel developed different algorithms for producing pronunciation variants depending on language and speaking style of the speakers [9]. M. Wester suggested pronunciation models which use both based on knowledge and data-driven.

Knowledge based methods use phonological linguistic rules which generally cannot capture the irregularities in the spontaneous speech. There is a gap between the linguistic knowledge found in the literature and the variations generated in the spontaneous speech. H. Strik and C. Cucchiaroni surveyed the literature covering various methods for modeling pronunciation variation [10].

## 3. THEORETICAL BACKGROUND

Figure 1 shows the schematic view of the vocal mechanism in humans. Vocal tract is one of the main articulators. It connects vocal cords to lips and consists of pharynx and mouth. The pharynx is the connection between esophagus to the mouth. The total length of the vocal tract in a male is 17 cm. The cross sectional area varies from zero when is completely closed to a maximum 20 square cm. Tract between velum and nostrils is called nasal tract. It produces the nasal sounds when the velum is lowered and the nasal cavity is acoustically connected to the vocal tract.

When the human takes breath, air enters the lungs. When the air escapes, the vocal cords are caused to vibrate. Articulators like jaws, tongue, mouth, lips and velum adjust their positions and produce the desired sounds.

Linguistically distinct speech sounds are called phonemes. A set of articulators are used to generate a phonetic sound. When the human being speaks a word, the articulators change their positions temporally to generate a sequence of phonetic sounds. The articulators are the vocal cords, pharyngeal cavity, velum, tongue, teeth, mouth, nostrils, etc. The articulators and the positions they assume while generating a phoneme are called features corresponding to that phoneme.

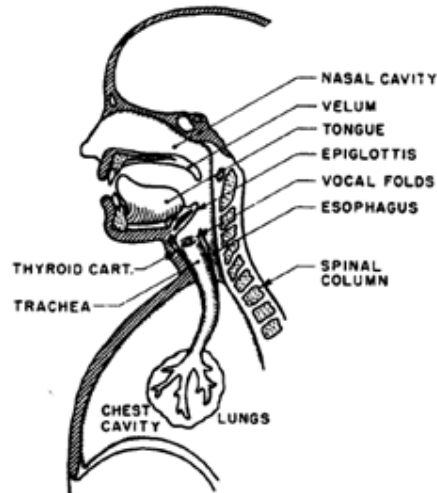


Figure1: The schematic diagram of vocal mechanism

The Standard English language has thirty-nine phonemes as shown in Figure 3.2. It consists of eleven vowel sounds, four diphthongs, four semi vowels, four nasal sounds, six stops, eight fricatives, two affricates and one whisper.

#### 4. DISTANCE BETWEEN VARIOUS PHONEMES

Distance between one phoneme to another is termed as phonetic distance between them. It is measured using the configuration of the articulators while generating the two phonemes [12, 13]. The positions assumed by the articulators while generating the phoneme sound are called its feature set.

The methodology followed for computation of distances between various pairs of phonemes is described in this section.

Table 1: Weightage assigned to features at various levels

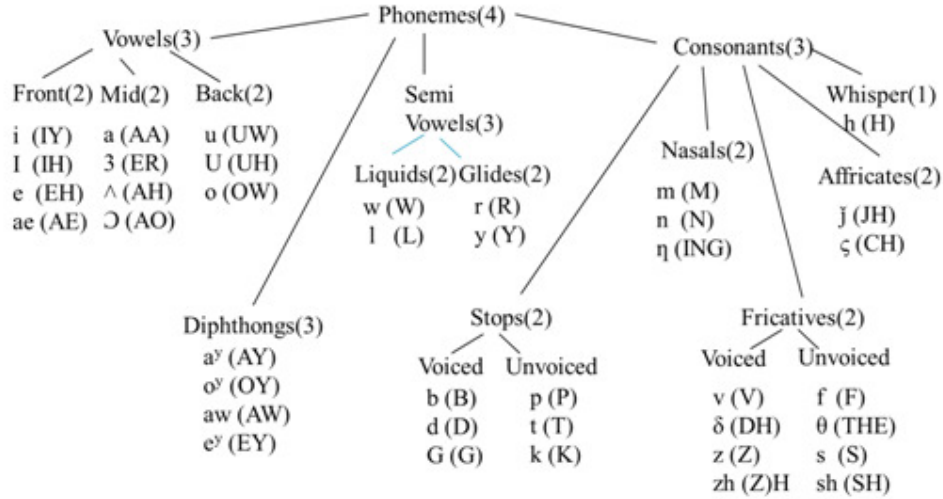
Level No.	Features	Weightage
1	Phoneme (root level)	4
2	Vowel, diphthong, semi-vowel, consonant	3
3	Front, mid, back, liquids, glides, nasals, stops, fricatives, affricates	2
4	All other features	1

Articulatory feature sets for various phonemes are extracted from the classification chart shown in figure 2. Features at various levels in the chart are assigned weightages as shown in table 1. The weights are extrapolated on the classification chart as shown in figure 2. The feature sets and their weightages for various phonemes are worked out.

The distance between two phonemes  $P_a$  and  $P_b$  is given by the Jaccard Coefficient

$$JC(Pa, Pb) = [1 - k * (Fa \cap Fb) / (Fa \cup Fb)] \quad (1)$$

K is a constant which is calculated experimentally.



Note: The figures within the brackets indicate the weight assigned to the attached feature tag. Weight '1' is assigned to the tags where the figures are not indicated.

Figure 2: Classification of Standard English phonemes with weights assigned to various features

The computations are as under.

**Example:** Phonetic distance between a front vowel IY and a nasal M is computed as follows.

- Feature set Fa for the front vowel (Pa = IY) = {phoneme, vowel, front, high tense}.
- Feature set Fb for the nasal (Pb = M) = {phoneme, consonant, nasal, alveolar}.
- Features common to the feature sets Fa and Fb = (Fa ∩ Fb) = {Phoneme}
- Weightage of the features common to both the feature sets W (Fa ∩ Fb) = 4.
- Total features in both feature sets Fa and Fb = (Fa ∪ Fb) = {phoneme, vowel, front, high tense, consonant, nasal, alveolar}.
- Weightage of total features in both the feature sets W ((Fa) ∪ (Fb)) = {4 + 3 + 2 + 1 + 3 + 2 + 1} = 16.
- Jaccard Similarity Coefficient JC (Pa, Pb) = W (Fa ∩ Fb) / W (Fa ∪ Fb) = 4 / 16 = 0.25.
- Jaccard Distance JD (Pa, Pb) = 1 – JC = 0.75.

### 3.2 Phoneme Substitution Cost Matrix

The substitution cost is the cost for replacing one phoneme with the other. The phonetic distances between 1521 pairs of phonemes are estimated.

### 3.3 Edit Operations

The three edit operations are substitution, insertion and deletion operations. Half of the substitution cost is taken as one Indel.

## 5. DPW ALGORITHM

DPW algorithm uses dynamic programming for global alignment. Needleman-Wunsch algorithm is modified to suit the usage of the algorithm in DPW algorithm. The phoneme cost matrix is used in place of similarity matrix and the Indel is used in place of the gap penalty. All the cells in the similarity matrix are filled using the substitution, and indel values. Bottom right hand corner cell value is phonetic distance between the given sequences.

Flow chart for DPW algorithm is given in figure 3.

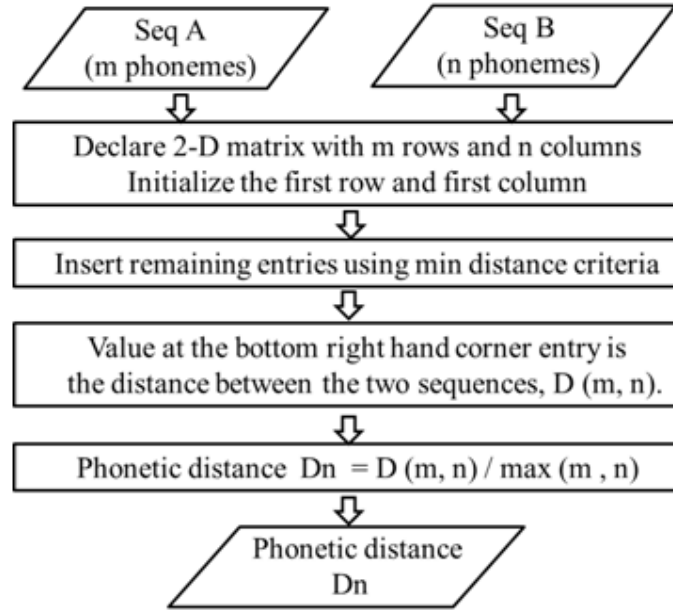


Figure 3: Flow chart for DPW algorithm

## 6. EXPERIMENTATION AND RESULT ANALYSIS

In this section, computation of phonetic distances using DPW algorithm is illustrated. The pronunciations and words are represented by their phoneme sequences.

### Data Source

Datasets are drawn from CMU pronunciation dictionary (CMUDICT). The CMUDICT has 130984 orthographic words followed by its phoneme sequences, out of which 8513 words have multiple pronunciation phoneme sequences.

### Experimental Setup

The experimental setup to measure the phonetic distances using DPW algorithm is shown in figure 4.

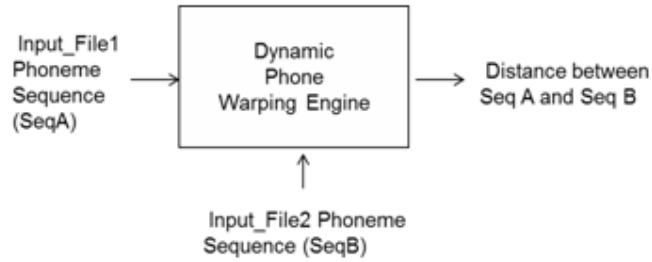


Figure 4: Test setup to compute phonetic distance using DPW algorithm

One sequence of phonemes are listed in file1 and the another sequence is taken File2. The algorithm described in flow chart is used to compute distance between the two sequences.

## 6.1 RESULT ANALYSIS

**Experiment 1:** Computation of phonetic distance in five different test cases is experimented and the results are recorded table 4.1.

The summary of normalized phonetic distances for test cases 1 to 5 is shown in table 2.

Table 2: Summary of normalized phonetic distances for test cases 1 to 5

Test Case No.	Details	Normalized Phonetic Distance
1	Same word, Same pronunciation compared with itself	0
2	Same word with different pronunciations	0.062
3	Same word with different pronunciations of unequal length	0.14
4	Different words with an unequal number of phoneme sequences	0.485
5	Different words with an equal length of phoneme sequences	0.358

In table 4.1, inter-pronunciation distances are computed in test cases 2 and 3 and inter-word distances are computed in test cases 4 and 5. It may be noted that the inter-word distances are greater than half of the Indel value and inter-pronunciation distances are less than half of the Indel value.

The results from the above five test cases reveal that the inter-pronunciation phonetic distance is less than inter-word phonetic distance.

The hypothesis resulted from the experimental results is that the distance between two pronunciations ( $D_A$ ) of a word is significantly less than phonetic distance between any two words ( $D_w$ ). It is possible to classify that a given sequence of phonemes is pronunciation variant or a new word itself.

$$D_A < D_w \quad (2)$$



Statistical z statistic tests have validated and confirmed the above hypothesis.

**Experiment 2:** 109 pairs of phoneme sequences and average phonetic distance calculations.

**Criterion for Error Count** Let the threshold phonetic distance for classification is half the value of one Indel. Let us call it as Critical Distance (Dc). The criterion for counting errors is as follows:

Let Wa and Wb be serial numbers of the words in test\_file1 and test\_file2 respectively. An error is counted in case the normalized phonetic distance (Dn) between a pair of pronunciations (Word A (Wa) = Word B (Wb)) is greater than Dc or the Dn for a pair of different words is less than or equal to Dc.

If  $\{((Dn > Dc) \ \&\& \ (Wa = Wb))\} \parallel \{(Dn \leq Dc) \ \&\& \ (Dn \neq Wb)\}$

Increment Error\_count; (3)

## Results

Result summary is shown in Table 3.

Table 3: Results of comparison of 109 pairs of words

Total Number of input word pairs analyzed	= 109
Total Number of errors	= 12
Classification Error Rate	= 11.01%

## Analysis

Experiment 2 gives the DPW results of 109 pairs of phoneme sequences corresponding to 55 different words with different lengths of phoneme sequences. The results show that the average normalized distance between the any two pronunciations is 0.069 and the average phonetic distance between the any two different words is 0.247. These results support that the inter-pronunciation phonetic distance is less than inter-word phonetic distance. Further experiments are carried out with larger datasets and hypothesis testing is carried out using z test statistic.

## 7. APPLICATION OF THE MODEL

The DPW algorithm is generic and can be used to classify a given sequence of phonemes corresponds to a new word or a pronunciation variant of an existing word in the dictionary. A critical distance threshold criterion can be developed to classify the given utterance into pronunciation variant or new words.

The phonemic distance measurements using DPW algorithm is independent of any particular language. Basically, it is using the phoneme set of that particular language. It can be used for any language generically. For instance, to utilise this algorithm for an Indian language, say Telugu, the phoneme set of Telugu language will be used in place the phoneme of English language.

Well-known languages like English have well-developed pronunciation dictionaries. But there are many sparse languages which are less known and do not have the readily available pronunciation dictionaries. The DPW algorithm can be used to build the pronunciation dictionaries for the sparse languages.

Pronunciation of a word differs from person to person. The DPW technology can be used to identify a speaker based on the pronunciation style. Therefore, the DPW algorithm can be used for speaker recognition.

A human being can understand and register pronunciation variability. But the machines like Interactive Voice Response (IVR) systems need supervised training to do so. The DPW technology can be used to build the online pronunciation capability in IVR systems. This the DPW technology has speech recognition applications.

The Information Technology (IT) companies have customers from all over the globe. The employees of the foreign companies will have the accent of their native language. The employees of Indian IT companies are educated to understand the accent of their native language. There is a challenge for the Indian employees understand the accent of foreign customers during initial stages. A pronunciation translator can be built to help the Indian employees to get over the above problem.

## 8. CONCLUSIONS

In this paper, the generation and perception of human speech is described. The phonemes are classified based on the articulatory features. The articulatory feature sets for generation of phonetic sounds are worked out. Weightage is assigned to each feature in the feature set and the total weightage of the feature set for each phoneme is computed. Phonetic distances between various pairs of the phonemes of the Standard English language are computed.

DPW algorithm is described with the help of a flow chart and is illustrated with the help of test cases. The analysis of the results led to the formulation of a hypothesis which gives the relationship between the inter-pronunciation distance and the inter-word distance. The hypothesis is tested at 1% significance level using z-test statistic.

## REFERENCES

- [1] Samuel Silva, António Teixeira, Unsupervised segmentation of the vocal tract from real-time MRI sequences, *Computer Speech & Language*, Vol 31, Volume 33, Issue 1, Pages 25-46, 2015.
- [2] S.-A. Selouani, Y. Alotaibi, W. Cichocki, S. Gharsellaoui, K. Kadi, Native and non-native class discrimination using speech rhythm- and auditory-based cues, *Computer Speech & Language* Volume 31, Issue 1, Pages 28-48, 2015.
- [3] Mahesh Kumar Nandwana, Ali Ziaei, John H. L. Hansen, Robust unsupervised detection of human screams in noisy acoustic environments, *IEEE Proceedings on Audio, Speech and Signal Processing, ICASSP 2015*, 161 – 165, 2015.
- [4] B. Yegnanarayana, S. Rajendran, Hussien Seid Worku and N. Dhananjaya, Analysis of glottal stops in speech signals, *IEEE Proceedings, INTERSPEECH 2008*, Brisbane, Australia, pp. 1481-1484, Sep. 22-26, 2008.
- [5] Jennifer E. Arnold, Michael K. Tanenhaus, Disfluency effects in comprehension: how new information can become accessible, In Gibson, E., and Perlmutter, N. (Eds) *The processing and acquisition of reference*, MIT Press, JANUARY 2011, pp 1-30.
- [6] Baker, J. M., Li Deng, Sanjeev Khudanpur, Chin-Hui Lee, James Glass and Nelson Morgan. 2009. Historical Developments and future directions speech recognition and understanding. *IEEE Signal Processing Magazine*, Vol 26, no. 4 78-85, Jul 2009.

- [7] Stefan Hahn, Paul Vozila, Maximilian Bisani, Comparison of Grapheme-to-Phoneme Methods on Large Pronunciation Dictionaries and LVCSR Tasks, IEEE proceedings of INTERSPEECH 2012.
- [8] M. Divay and A.-J. Vitale. Algorithms for grapheme-phoneme translation for English and French: Applications for database searches and speech synthesis. Computational linguistics, 23(4):495–523, 1997.
- [9] M. Adda-Decker and L. Lamel. Pronunciation variants across system configuration, language and speaking style. Speech Communication, 29:83–98, 1999.
- [10] M. Wester. Pronunciation modeling for ASR- knowledge-based and data-driven methods. Computer Speech and Language, pages 69–85, 2003.
- [11] H. Strik and C. Cucchiaroni. Modeling pronunciation variation for ASR: A survey of the literature, Speech Communication, 29(4) (1999) 225–246.
- [12] L. Rabiner, B. Juang and B. Yegnanarayana, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, N.J., 2010.
- [13] Amos Tversky, Features of Similarity, Psychological Review. Vol 84, Number 4, July 1977.

## AUTHORS

**Akella Amarendra Babu** received B. Tech (ECE) degree from JNU, M. Tech (CSE) degree from IIT Madras, Chennai and Ph. D. degree in Computer Science and Engineering from JNTUA, Ananthapuramu. He served Indian Army for 23 years as Lt Colonel in Corps of Signals and has 12 years of senior project management experience in corporate IT industry. He has two and half years research experience on mega defense projects in DLRL, DRDO and is working as Professor of CSE department in Engineering Colleges at Hyderabad. He published more than 20 research papers in various national and international conferences and journals. He published a book chapter and has a patent. His research interests include speech processing, computer networking, information security and telecommunications. He is a Fellow of IETE, life member of CSI and IAENG.



**Y Rama Devi** received B.E. from Osmania University in 1991 and M. Tech (CSE) degree from JNT University in 1997. She received her Ph. D. degree from Central University, Hyderabad in 2009. She is Professor, Chaitanya Bharathi Institute of Technology, Hyderabad. Her research interests include Speech and Image Processing, Soft Computing, Data Mining, and Bio-Informatics. She is a member for IEEE, ISTE, IETE, IAENG and IE. She has published more than 50 research publications in various national, international conferences, proceedings and journals.



*INTENTIONAL BLANK*

# A SURVEY ON QUESTION ANSWERING SYSTEMS: THE ADVANCES OF FUZZY LOGIC

Eman Mohamed Nabil Alkholy<sup>1</sup>, Mohamed Hassan Haggag<sup>1</sup> and  
Constantine's Koutsojannis<sup>2</sup>

<sup>1</sup>Faculty of Computers & Information, Helwan University, Cairo, Egypt

<sup>2</sup>Health Physics & Computational Intelligence Lab, TEI of Western Greece,  
Aigion

## ABSTRACT

*In real world computing environment with using a computer to answer questions has been a human dream since the beginning of the digital era, Question-answering systems are referred to as intelligent systems, that can be used to provide responses for the questions being asked by the user based on certain facts or rules stored in the knowledge base it can generate answers of questions asked in natural , and the first main idea of fuzzy logic was to working on the problem of computer understanding of natural language, so this survey paper provides an overview on what Question-Answering is and its system architecture and the possible relationship and different with fuzzy logic, as well as the previous related research with respect to approaches that were followed. At the end, the survey provides an analytical discussion of the proposed QA models, along or combined with fuzzy logic and their main contributions and limitations.*

## KEYWORDS

*Question-answering, Neutral language processing, Fuzzy Logic, Answer Extraction, Evaluation Metrics*

## 1. INTRODUCTION

As technology and human-computer interaction advances, there is an increased interest in affective computing and a large amount of data is generated and made it available every day, and for that it requires to integrate and query a huge amount of heterogeneous data, for that NLP has been recognized as a possible solution that capable to manipulate and represent the complex query as uncertain and complicated that are existing in them. Which leads to the generation of QA consisting of Equation and Answer that mapping between these information [1]. However, Question Answering (QA) is a fast-growing research area that combines the research from Information Retrieval, Information Extraction and NLP. It can be seen as the next step in information retrieval is to automatically generating answers to natural language questions from humans, that allow users to pose questions in natural language and receive succinct answers.

In other aspect (view) Meaning is essential to human language. The idea of the Fuzzy Is use fuzzy set theory that it derived from it a form of multivalued logic to deal with approximate reasoning, to represent and process the linguistic information and attributes of the real world [1]. This survey provides an overview of QA system its system architecture and the possible relationship and differentiation with fuzzy logic, our believe is to have human-friendly dealing with more accuracy in information that needs to use fuzzy logic with NLP.

## 2. BACKGROUND

Over the past four decades Question Answering systems have been transitional much at par with the whole of natural language processing. In this section, we present a previous work on development of QA system and it's propose, the earliest system was developed in 1959 (in the spirit of the era called The Conversation Machine), and A large number of QA system have been developed since 1960's One of the most memorable systems was BASEBALL developed by (Green et al. 1961 in NL DB systems) [2]. Although, capable of answering rather complex questions, BASEBALL was, not surprisingly, for answering questions about baseball games played in the American league over one season, restricted to questions about baseball facts, and most question answering systems were for a long time restricted to front-ends to structured databases. And in 1963 they develop QA system PROSYNTHES that permit user to ask a question in English, it accept statements in (sub set of English) as input to its database and accepts quotations as a query to the database. And for read and solve the kind of word problems they develop (Problem-solving systems) STUDENT QAs system in 1964 That can read and solve the kind of word problems found in high school algebra books. The program is often cited as an early accomplishment of AI in natural language processing [3].

In early 1966 they provide QA system ELIZA that can communication with user) And This conversation can be simulated with a psychologist. It was able to converse on any topic by resorting to very simple rules that detected important words in the person's input, And in 1972 they develop SHRDLU that offered the possibility to operation of a robot in a toy world (the "blocks world)to be simulated with ability to ask the robot questions about the state of the world SCHOLAR QA system develop in 1973 it was a formal theory characterizing the variety of plausible inferences people use to ask questions about which their knowledge is incomplete. it has methods embed to lessons learned from such research into the SCHOLAR system [4].

In the same year 1973 they develop the first automatic question answering system (SAM). After three years in 1976 TRIPSYS (HWIM) was the first QA that understand speech Question , TRIPSYS(HWIM) was developed as the context for a research project in continuous speech understanding- it understands and answers questions about planned and taken trips, travel budgets and their status, costs of various modes of transportation to various places it's called HWIM (for "Hear What I Mean[2]). And the second famous QA system was (NL DB systems) Problem solving systems LUNAR in 1977[5] for answering questions about the geological analysis of rocks returned by the Apollo moon missions.

In 1977 develop two QA system the first one is GUS it was a dialog system for airline reservation second one was LIFER that develop to asking questions about U.S. Navy ships [6]. This system used a semantic grammar with domain information built within in 1978 they start to develop system that deal with story comprehension (NL DB systems) QUALM an application that use for story comprehension and this application responsible for scripts and plans in a very restrictive

domain. In 1983 Salton and McGill describe question answering (QA) systems as mainly provide direct answers to questions. Finally Kupiec (1993) employed similar but rather simpler WH question models to build a QA system [8].

In 1991 was QA system (LILOG) text understanding system that operated on the domain of tourism information in a German city. In 1993 they start to combined NLP with the use of an online encyclopedia by developing MURAX QA system that combined NLP with the use of an online encyclopedia with hand-coded annotations to sources. In subsequent developments, QAS aimed on making linguistic analysis of the questions to capture the intended requirements in a natural way [7]. In subsequent developments, one of QAS aimed to capture the intended requirements in a natural way IS to making linguistic analysis of the questions.

In recent 90's, question answering achieved a great progress due to the introduction of the Text Retrieval Conferences (TREC) question answering track there has been great progress in open domain Question answering (Voorhees 2001) . These systems use unrestricted text as a primary source of knowledge. One such system, MASQUE (ANDROUTSOPOULOS et al., 1993) use logic representation to represents natural language questions, and for retrieving intended information from database, it use to translates the logic query into a database query. It separates from mapping process the task of linguistic process. FAQ Finder (Burke et al., 1997) does matching of the question list that compiled in a knowledge base with questions through statistical similarity and semantic similarity and for syntax-based natural language understanding technique. In 1999 was LASSO that win the question answering task. It used question classification technique and syntaxbased natural language understanding technique. Another QAS in (2000) was developed by Riloff and Thelen (QUARC), that has the ability to classify questions into different types and use lexical and semantic clue to derive their expected answer [9] [10].

Later, the focus of developing QASs was shifted toward open domain QAS, TREC Evaluation campaign which is taking place every year since 1999 to manage and query large volume of data and represent most of research in open domain question answering from unstructured data sources, and that lead to question-answering evaluations as a recent success started as part of the Text Retrieval Conference (TREC). The best systems are now able to answer more than two thirds of factual questions in this evaluation, it describes the results and the associated evaluation Methodology develop by Ellen Voorhees, the second paper, by Buchholz and Daelemans, explores the requirements for answering complex questions that have compound answers or multiple correct answers [11].

The third paper, by Lin and Pantel, describes a new algorithm to capture paraphrases that allow a more accurate mapping from questions to potential answers. The fourth paper, by Light, Mann, Rilo and Breck, describes experiments that systematically factor and assess question answering into component sub-problems. The first TREC evaluation campaign provides a list of 200 questions and a document collection. The answers were known to be present in the collections. The maximum lengths of answers were allowed to be 50 or 250 characters[12]. Systems were asked to give 5 ranked lists of answers.

In the next campaign, TREC-9 held in 2000, the number of questions and size of document collections were increased. In TREC-10 in 2001, a new complexity with respect to answers. The lengths of answers were reduced to 50 words [11]. In TREC- 11, held in 2002, systems were expected to give exact short answers to the questions. In TREC from 2002 to 2007, the list of questions, definition questions, and factoid questions were included in the evaluation campaigns.

In TREC 2005 and TREC 2006 there was temporal question in addition to 75 topics that contains different type of Questions. In TREC 2007, there were a collection of documents that include collection of blogs, and progress competitions that lead to increasing documents collection complexity, and questions complexity, and that also effect to increasing answer evaluation strategies complexity.

In 2000 they start to develop QA system for Unix operating system Unix Consultant For answered questions pertaining to the Unix operating system [10]. This system had ability to accommodate various types of users by phrasing the answer, and it has its domain that contain a comprehensive hand-crafted knowledge base. In 2001 there was two QA systems the first one was INSIGHT question answering system which uses some surface patterns, wins the question answering task in TREC-10, and the second one was SiteQ use the density-based extraction method to retrieve related passages first and then extract the exact answers in them, which can greatly improve the extraction speed.

In 2002 STARTQA system was the first web-based QA system for English and in the same year 2002 was Answer bus QA system(ODQA) they develop this system to accepts questions in several languages (extend the answer extraction process from the local data source to the World Wide Web, which allows them to deal with large count of questions. After one year they develop QA system (ARANE) it was first fully downloadable open-source Web-based factoid question answering system, and in the same year 2003 create a QA system AQUA was a sophisticated automatic question answering system, which combines natural language understanding technique, ontological knowledge, logical reasoning abilities and advanced knowledge extraction techniques [12].

In 2008 they start to develop a new QA system with different approach and it was a List questions ask for different instances of a particular kind of information to be returned. In 2012 they develop QA Systems that oriented to work with opinions rather than facts can also be included in this category. In 2014 was QAKIS by Cabrio et al and FREITAS14 develop by Freitas and Curry and also INTUI3 by Dima, in 2015 was develop three QA system first was HAKIMOV15 by Usbeck et al [12]. and second was QASYO by Hakimov et al. last. However, QA systems have developed over the past few decades until they reached the structure that we have nowadays.

## **2.1 Implementation Approaches for QA Systems**

The basic aim of QA System is to provide correct and short answer with high accuracy to the user and There are many Approaches used in Question answering system based on different purpose [13]. This section will present an implementation approaches for various categories of QA System such as:

### **2.1.1 First Approach**

That relied on artificial intelligence (AI) it called Linguistic approach and it has the ability to build QA logics by using methods that integrated natural language processing (NLP) techniques and knowledge base. It used Linguistic techniques for formulating user's question into a precise query that merely extracts the respective response from the structured database, it Used to understand natural language text, linguistic & common knowledge Linguistic techniques such as such as (tokenization, POS tagging and parsing [14].



### 2.1.2 Second Approach

(rule-based approach) that rely on the (rule-based mechanism), they are built to identify question classification features. Quarc developed by Rilloff et al., and Cqarc [15] developed by Hao et al. Used to identify the question class by using semantic clues in question and for looking lexical they used heuristic rules.

### 2.1.3 The Third Approach

(Statistical approach) gives the better results than other approaches user can use this approaches successfully to applied the different stages of a QA system, it independent of structured query languages and can formulate queries in natural language form. IBM's statistical QA system was based on the statistical model. This system utilized maximum entropy model for question/ answer classification based on various N-gram or bag of words features. Moschitti had used Rocchio and SVM text classifiers for question and answer categorization. Availability of huge amount of data on internet increased the importance of statistical approaches [16].

### 2.1.4 The Fourth Approach

is Pattern matching that have the ability to replace the sophisticated processing involved in other competing approaches by using the expressive power of text patterns. Some of patterns matching QA rely on templates for response generation while most of of the patterns matching QA systems [17].

### 2.1.5 The Fifth Approach

is the Surface Pattern based approach. This approach rely on an extensive list of patterns to extracts answers from the surface structure of the retrieved documents. It's automatically learning based pattern or it is human crafted. [18]. However, to identify any question answer it depend on the basis of similarity between their reflecting patterns having certain semantics as well as their Methods, definition and Characteristics, Limitation and Aims for each QA system that was developed before which will be helpful for new directions of research in this area (Table 1).

Table 1: List of Most Popular QA Systems

QA system	Method & definition	Characteristics and Implementation Issues	Limitation
BASEBALL [2]	<b>Description:</b> answers English questions-about-the scores, teams, locations, and dates of baseball games. <b>Method:</b> Analyzed the question, using-linguistic knowledge-, into canonical form .	<b>Characteristics:</b> Input sentences have to be simple, and not contain-sentential- connectives, such as (and, or). - The data are stored in a data base in attribute-value format. - questions-transformed into the same format, but in an automatic way. <b>Implementation:</b> system was implemented in the Lincoln Laboratory. And it relied on Linguistic approach.	- its domain of {baseball only} (Close domain) -the database tied to specific domains, where the attribute-value structures can be uniform, and the types of questions are limited.
STUDENT [3]	<b>Description:</b> Correctly solving the algebra problem <b>Method:</b> was taken-as-a demonstration-that the system understood the written statement of the problem.	<b>characteristics:</b> Understanding-language-requires-world-knowledge. -read and solved high school algebra word problems. <b>Implementation :</b> It is written in Lisp	- systems were limited by the amount of knowledge-they contained. Closed Domain

ELIZA [4]	<p><b>Description:</b> system that simulated a Conversation with a psychologist.</p> <p><b>Method:</b> the computer can read messages typed on the typewriter and respond by writing on the same instrument.</p>	<p><b>Characteristics:</b> ability to converse on any topic by resorting to very simple rules that detected important words in the person's input.</p> <p><b>Implementation:</b> it was first implemented in the SLIP language, that use as an extension to FORTRAN but with better functionality to process doubly linked lists, and Its present implementation is on the MAC time-sharing system at MIT. And relied on Linguistic approach.</p>	<p>- (closed domain) the knowledge stored in the structured database was only capable of answering questions asked within the restricted domain.</p>
LUNAR [5]	<p><b>Description:</b> Designed to enable a lunar geologist to conveniently access the chemical analysis data on lunar rock that was accumulating as a result of the Apollo moon mission".</p> <p><b>Method:</b> The entries in the analysis table-specify-the concentration of some constituent in some phase of some sample.</p>	<p><b>characteristics:</b> -The system contains two data bases: a 13,000-entry table off chemical and age analyses of the Apollo 11 samples. - able to answer 90% of the in-domain questions posed by working geologists, without prior instructions as to phrasing.</p> <p><b>implementation</b> system is implemented in LISP. Parse English question into a data base query, Syntactic analysis via augmented transition network parser and heuristics and relied on Linguistic approach. That can analyze automatically with a transition network parser and translated into a data base query language.</p>	<p>-Sophisticated, with the syntax and semantics of questions - having a particular database could not be easily modified to be used with different databases. -closed domain</p>
GUS [6]	<p><b>Description :</b> Simulated a travel advisor.</p> <p><b>Method:</b> it attempt to explore the integration of already existing programming technology for a performance demonstration.</p>	<p><b>characteristics:</b> had access to a restricted database of information about airline flights. - is a steady genetic algorithm with subpopulation support.</p> <p><b>Implementation:</b> system implemented in MChART ,it is a framework for implementing parsers. And relied on Linguistic approach.</p>	<p>-closed domain -the knowledge stored in the structured database was only capable of answering questions asked within the restricted domain.</p>
MURAX [7]	<p><b>Description:</b> searched the encyclopedia for noun Phrases identified from the question.</p> <p><b>Method:</b> exploited the phrase relations contained in the question to match questions with answer hypotheses.</p>	<p><b>Characteristics:</b> -used the technology of robust shallow parsing. -use the Internet as a corpus - answers hypothesize noun phrases</p> <p><b>Implementation:</b> the sentences of matching text are selected to confirm phrase relations implied by the question, rather than being selected solely on the basis of word frequency.</p>	<p>-lack of basic-information extraction support -open domain</p>
QUALM [8]	<p><b>Description:</b> designed to test a story understanding system.</p> <p><b>Method:</b> -it requires question to be parsed and a conceptual graph to be built, and it include two stages: understanding the question and finding the answer.</p>	<p><b>Characteristics:</b> it able to answer questions about ideas not specifically mentioned in the texts. - represented as a conceptual graph and text comprehension</p> <p><b>Implementation:</b> - it implemented as a language-independent question answering module, and integrated into other natural language processing applications</p>	<p>-hard to classify and complex system - (closed domain) highly restricted in terms of the domain, or world, knowledge required and the genre of text covered</p>
SHRDLU [9]	<p><b>Description:</b> it's a research system that help researchers understand the issues involved in modeling human dialogue.</p> <p><b>Method:</b> the system answers questions, execute commands, and accepts information in normal English Dialog.</p>	<p><b>characteristics:</b> contain a combination of syntax, semantics, and reasoning. - it uses semantic information and context to understand discourse and to disambiguate sentences. -it could search back further through the interactions.</p> <p><b>implementation :</b> its written in Micro Planner and Lisp ( a general-purpose, multi-paradigm programming language.</p>	<p>-constrained to a simple block world. -restricted-domain - has a core database handwritten by experts</p>
QAKIS [10]	<p><b>Description</b> allows end users to submit a query to an RDF triple store in English and obtain the answer in the same language.</p> <p><b>Method:</b> it composed of two main modules the <b>query generator</b> takes the user question as input, generates the typed questions, and then generates the SPARQL queries from the retrieved patterns.</p>	<p><b>Characteristics:</b> huge amounts of available semantic data -implement a relation-based match for question interpretation, to convert the user question into a query language</p> <p><b>Implementation:</b> QAKIS addresses the task of QA over structured Knowledge Bases (KBs) where the relevant information is expressed in unstructured form to implement a relation-based match for question interpretation, to convert the user question into a query language.</p>	<p>Open domain</p>



IBM's statistical QA system [16]	<b>Description:</b> it's IBM's static question answer for TRCE 9. Is an application of maximum entropy classification for QA predication and name entity marking <b>Method:</b> maximum entropy	<b>characteristics:</b> classification based on various N-gram or bag of words features. - Trained on crops of DSTS--use REASON <b>Implementation:</b> Maximum Entropy Model for this (IBM's QA) system used a two-pass approach based on Okapi formula and expansion of queries based on TREC-9 QA corpus. -Statistical approach based	Open domain
A WebBased Automatic QAs [19]	<b>Description:</b> presents an answer extraction method which based on the calculation of sentence similarity between question and answer. (Sentence Similarity Model)	<b>characteristics:</b> using Web data resource as database for question answering system - use answer correctness as our evaluation me <b>Implementation:</b> Statistical models applied for question classification	-Open domain -difficulty of finding answers to questions little data available in the target language in which to search for answers.
NLDB [20]	<b>Description:</b> it's a database system Engaged the user in dialogues ,it allows the user to access information stored in a database.	<b>characteristics:</b> used with large databases. -could be configured to interface to different underlying database. <b>Implementation:</b> it was implemented entirely in Prolog. It transformed English questions into Prolog expressions, which were evaluated against the Prolog database. The code of Chat-80 was circulated widely)	Closed Domain the knowledge stored in the structured database was only capable of answering questions in restricted domain.
Statistical Approaches to Answer-Finding [21]	<b>Description:</b> investigates whether a machine can automatically learn the task of finding, within a large collection of candidate responses <b>Method:</b> use four statistical techniques for answer-finding	<b>characteristics:</b> Usenet FAQ documents and customer service call-center dialogues from a large retail company <b>Implementation</b> - statistical models applied for question classification. using , adaptive tf. idf algorithm and automatic query expansion with latent variable model.	Open domain
PHLIQA1 [22]	<b>Description:</b> it Designed to answer short questions against a data base containing fictitious data about computer installations in Europe and companies using them. <b>Method:</b> - use of the lambda calculus	<b>Characteristics:</b> Have three division of the translation from natural language questions to data base 1-English-oriented Formal Language (EFL). 2-World Model Language (WML). 3-Data Base Language (DBL). <b>Implementation:</b> in PHLIQA1 system Questions are translated into a formal language used to access the data base.	the limitation to a narrow domain
(LADDER) [23]	<b>Description:</b> it's a database systems It was designed as a natural language interface to a database of information about US Navy ships. <b>Method:</b> used semantic grammars to parse questions to query distributed database. it candevloped as a prototype system for understanding questions posed in English about a naval domain.	<b>Characteristics:</b> used with large databases . - grammar must be tailor-made for each given database. - only support simple one table queries or multiple table queries . <b>Implementation:</b> used a semantic grammar to parse questions and query a distributed database.it translated each English question into one or more relational database queries, prosecuted the queries on a remote computer.Based on LIFER parser, which interpreted sentences according to a 'semantic grammar.'	-a different grammar had to be developed whenever Ladder was configured for a new application. - a narrow domain
Ask Jeeves [24]	<b>Description :</b> allowed end-users to teach the system new words and concepts at any point during the interaction it's database systems <b>method:</b> The user stated requests in English, and Ask suitable requests to the appropriate underlying systems.	<b>Characteristics:</b> has its own built-in database. - use the Internet as a corpus <b>Implementation :</b> its pointing the questioner to Web links that might contain information relevant to the answer to the question.	All the applications connected to Ask were accessible to the end-user through natural language requests -open domain

Answer Bus [25]	<b>Description:</b> is an question answering system based on sentence level Web information retrieval <b>method :</b> Answer Bus extracts sentences that are determined to contain answers.	<b>Characteristics:</b> High performance & accuracy for response time.-use the Internet as a corpus. -It accepts questions in several languages. <b>Implementation:</b> is based on sentence level information retrieval. It accepts users' natural-language questions in different language and extracts possible answers from the Web.	-it only extracts the named entities that match question types -open-domain
AQUA [26]	<b>Description:</b> allow users to ask questions in everyday language and receive an answer quickly and with a context <b>Method:</b> Use a similarity algorithm to map between names of relations in the knowledge base and names of relations in the ontology	<b>Characteristics:</b> it use semantic annotations to perform inferences . - AQUA's inference engine operates within the framework of multi-sorted logic, <b>Implementation:</b> it uses NLP technology, Logic and a hand-crafted ontology. And using Dice coefficient and WordNet algorithm. This algorithm is used to ensure that the question does not fail .	Closed domain
START [27]	<b>Description:</b> it's a First Web-Based Question Answering System, Designed To Answer Questions That Are Posed To-It-In-Natural Language. <b>method:</b> It Parses Incoming Questions, Matches the Queries Created from The Parse Trees Against Its Knowledge Base And Presents-The Appropriate-Information Segments To The User.	<b>Characteristics</b> Ability to Answer millions of English-Questions-- use the Internet as a corpus - used technique called "natural language annotation" <b>Implementation</b> consists of two modules. The understanding module: analyzes English text and produces a knowledge base that encodes information found in the text. -the generating module: produces English sentences. Used in conjunction with the technique of natural language annotation. -Linguistic based QA systems	-ambiguous modification linguistically-uninformed QA - have difficulty handling-semantic symmetry - open domain

Table2: Fuzzy Logic Implementation Approaches For QA Systems

QA System	Method & Definition
Quantitative Fuzzy Semantics (L. A. ZADEH)(1971)	the first paper on fuzzy sets it can construct fuzzy query languages for purposes of information retrieval, and, possibly, to implementation of fuzzy algorithms and programs. [37]
A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges L. A. Zadeh (1972)	represent a hedge as an operator, linguistic hedge such as very, more or less, much, essentially, slightly. [35]
PRUF- A Meaning Representation Language For Natural Languages” (1978)	PRUF provide a basis for question-answering and inference from fuzzy premises [40].
Test-Score- Semantics as A Basis For A Computational Approach (1986)	test-score semantics provides a framework for the representation of the meaning of dispositions.[42]
A Computational Approach to Fuzzy Quantifiers In Natural Languages (1983)	deal with fuzzy quantifiers (denote the collection of quantifiers in NLP whose representative elements are: much, not many, very many, notvery many) that are treated as fuzzy numbers[37]

Fuzzy Quantifiers: A Natural Language Technique for Data Fusion 2001	intended to formalize the notion of ‘linguistic adequacy’. then argue that the models of the theory are plausible from a linguistic perspective.[37]
Fuzzy Logic-Based Natural Language Processing and Its Application to Speech Recognition 2017	it create a system that can learn from a linguistic corpus. it use fuzzy semantic relations to represented by words and use such relations to process the word sequences generated by speech recognition systems. [32]
The Fuzzy Formal Concept Analysis (FFCA) 2003	proposed a fuzzy FCA-based approach for conceptual clustering for automatic generation of concept hierarchy from uncertainty information.[43]
A Hybrid Approach Using Ontology Similarity and Fuzzy Logic For Semantic Question Answering 2014	the user enters a source string as a question. the first objective of the machine is to syntactically analyze the text from the source. using fuzzifier approach (semantic fuzzy ontology) [44]
Improving translation memory fuzzy matching by paraphrasing 2015	present an innovative approach to match sentences having different words but the same meaning. presented a method that improves the fuzzy match of similar, but not identical sentences. [41]
Semantic parsing on freebase from question-answer pairs	train a semantic parser that scales up to Freebase.method : build a coarse mapping from phrases to predicates using a knowledge base and a large text corpus. [33]
FLINTSTONES: A fuzzy linguistic decision tools enhancement suite based on the 2-tuple linguistic model and extensions	Use fuzzy linguistic tools called Flintstones that proposed to solve linguistic decision making problems based on the 2-tuple linguistic model in order to validate the performance of the software suite with real datasets[34]
A fuzzy linguistic approach for human resource evaluation and selection in software projects 2015	present a fuzzy linguistic approach that utilizes 2-tuple fuzzy linguistic terms and supports the selection of suitable human resources based on their skills and the required skills for each project task.[35]
Fuzzy logic in natural language processing 2017	they outline how model of the meaning of basic constituents of natural language (nouns, adjectives, adverbs, verbs) has been elaborated in FNL[39]

### 3. QA SYSTEM OVERVIEW

Question answering is a process that understanding user natural language query and has the ability to provide a correct answer and extract it from retrieving relevant documents, data, or knowledge base .A question answering (QA) Question-answering systems are referred to as intelligent systems ,that can be used to provide responses for the questions being asked by the user based on certain facts or rules stored in the knowledge base it can generate answers of questions asked in natural [28].

#### 3.1 QA System Purpose and Types

The purpose of a QA system is to provide correct answers to user questions in both structured and non-structured collection of data. there are three main types for the QA Depending on the target domain and the way questions are answered it will present in table.

Table 4: Types of QA Systems

TYPES OF QA SYSTEMS		ADVANTAGE
CLOSED-DOMAIN	<p>Built for very specific domains and exploit expert knowledge.</p> <p>The first QA systems of this type were developed in the 1960s. Any data not present in the database is usually considered out-of-domain.</p> <p>Closed-domain question answering deals with questions under a specific domain, This type of QA is easier for the vocabulary is more predictable, and ontologies describing the domain are easier to construct</p> <p>-Two of the most cited closed-domain QA systems are BASEBALL<sup>[2]</sup> and LUNAR<sup>[5]</sup>.</p>	<p>Deals with very specific data which usually does not contain ambiguous terms and as a result can be processed more easily.</p>
OPEN DOMAIN	<p>Can be asked about virtually any topic and can theoretically extract the answer from any textual collection.</p> <p>- It can deals with unrestricted topics. Hence, questions may concern any subject. The corpus may consist of unstructured or structured texts.</p> <p>- Two of good example open domain QA systems are AnswerBus<sup>[25]</sup> and Askjeeve<sup>[24]</sup></p>	<p>An alternative to search engines, available on the Web. Instead of providing a list of relevant keywords user just asks a question.</p> <p>-Automatically build queries and retrieve relevant documents, extract answers from the retrieved text snippets and present them as a confidence-ranked list to the user.</p>

### 4. QA SYSTEM ARCHITECTURE & ALGORITHM

#### 4.1 System Architecture

As shown in figure (1) a QA system contain three main part beside other supplementary components: question classification (QUESTION PROCESSING), information retrieval (DOCUMENT PROCESSING) and answer extraction (ANSWER PROCESSING). The user writes a question using the user query interface. Then this query is used to extract all the possible



answers for the input question. The architecture of Question Answering system is as shown in Figure1

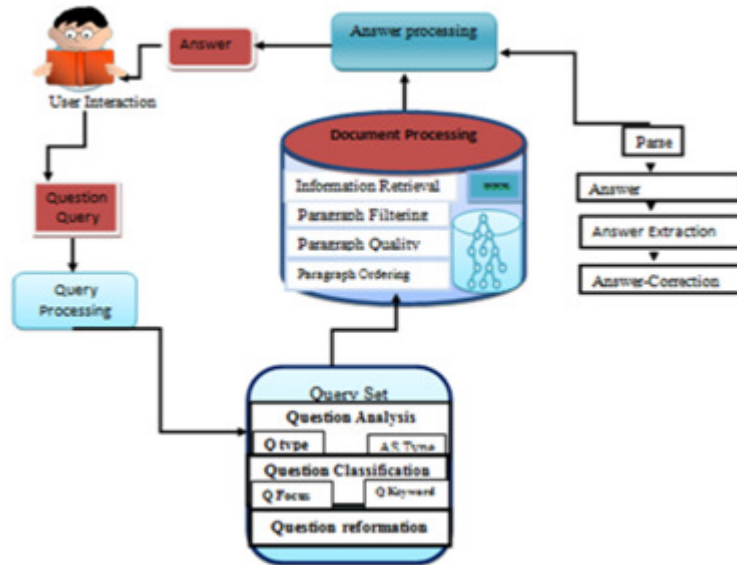


Figure (1): Question Answering System Architecture

## 4.1 Question Answering System Algorithm

### 4.1.1 Step One

(Question processing): Given a question as input, user writes his question by an interface the function of the question processing module is to process and analyses the question, and to create some representation of the information requested module is to process and analyses the question, and to create some representation of the information requested analyses the question, and to create some representation of the information requested. This leads to the classification of question that include three parts: Query Interface, question analyzer and Question classification. This step help to generate complex structured queries . and detects the expected answer type of a question e.g. the expected answer type of " When was last time that you feel this headache?" is date or "Do other family members have similar headaches?" Is yes/no this information helps guide the answer extraction process [29].

### 4.1.2 Step Two

after Input questions -Determining the question type: is selected from question taxonomy that system uses. and candidate answers are extracted using all the information gathered in the previous steps, e.g., keywords, entities and relations. The remaining candidate answers are then filtered and ranked according to different measures of similarity to filter out candidates with incompatible types. The Document Processing Module retrieves documents from the corpus that are likely to contain answers to the user's question. It consists of a query generation algorithm, text search engine and information retrieval that retrieve the relevant documents based upon important keywords appearing in the question. The query generation algorithm takes an input the

user's question and creates a query containing terms likely to appear in documents containing an answer. This query is passed to the text in system, which uses it to retrieve answer. [30].

#### 4.1.3 Step Three

The answer processing module is responsible for identifying, extracting and validating answers from the set of ordered paragraphs passed to it from the Document Processing Module, takes input from the retrieval component and tries to retrieve an exact phrase to return as an answer to achieve this required parsing and detailed question analysis by using of answer extraction algorithms. The identity answer extraction returns the CenterPoint of the passage, stripping words from either end until it fits within the specified answer. Then Answer Display The result and converted into required text which is required by the user and displayed to the user.

### 4.2 Evaluation Metrics

There are several parameters that are used to analyze the performance of different Question Answering Systems. In this section we describe some of the evaluation metrics used in Question Answering System to evaluate its performance such as: precision, recall and F-measure, a weighted harmonic mean of precision and recall, can be defined with respect to predicates for the purposes of QA evaluation. These precisions and recall metrics express true precision and recall, not approximations, when coupled with an answer key in which the judgments can be reasonably assumed to be exhaustive[31].

#### 4.2.1 Precision

Precision- Recall is the most common metric to evaluate information retrieval system. Precision is the ratio of retrieved documents that are relevant to all retrieved documents in the ranked list [30].

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (1)$$

#### 4.2.2 Recall

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad (2)$$

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

For example: text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query[36]. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents.



### 4.2.3 F-Measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

known as the  $F_1$  measure, because recall and precision are evenly weighted. It is a special case of the general  $F_\beta$  measure (for non-negative real values of  $\beta$ ):

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \quad (4)$$

Two other commonly used  $F$  measures are the  $F_2$  measure, which weights recall higher than precision, and the  $F_{0.5}$  measure, which puts more emphasis on precision than recall.

The F-measure was derived by van Rijsbergen (1979) so that  $F_\beta$  measures the effectiveness of retrieval with respect to a user who attaches  $\beta$  times as much importance to recall as precision"[36].

$$E = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$$

It is based on van Rijsbergen's effectiveness measure (5)

Their relationship is  $F_\beta = 1 - E$  where  $\alpha = \frac{1}{1 + \beta^2}$  (6)

## 5. FUZZY LOGIC OVERVIEW

This section presents a quick overview of what fuzzy logic is and its system architecture and the possible relationship with Question Answer, Fuzzy logic is a technique used for representing and manipulating uncertain information, In the more traditional propositional logic, each fact or proposition, the truth value in fuzzy logic may range between completely true and completely false that it's a concept of partial truth that fuzzy logic used to handle it. By contrast, in logic, non-numeric values are often used to facilitate the expression of rules and facts linguistic variable and that in fuzzy logic application such as age may accept values such as young and its antonym old. And Because natural languages do not always contain enough value terms to express a fuzzy value scale, it is common practice to modify linguistic values with adjectives or adverbs[37].

So any systems that cannot be precisely described by mathematical models or devices that have significant uncertainties or contradictory conditions, and linguistically controlled devices or systems can use fuzzy logic. As Lotfi Zadeh once stated, fuzzy logic is not going to replace conventional logic or methodologies, rather it will supplement them in circumstances where conventional approaches fail to solve a problem effectively [38]. In recent years, Fuzzy logic has proved to be particularly useful in expert system and other artificial intelligence applications and NLP. Therefore, many of the previous researchers in academia, applications include modeling, evaluation, optimization, decision making, control, diagnosis and information Measure the

growing interest in fuzzy logic. In particular, fuzzy logic is best suited for control-systems fields. it has been applied in areas such as breakdown prediction of nuclear reactors in Europe, earthquake forecasting in China, and subway control in Japan.

## 5.2. Fuzzy Logic General Architecture

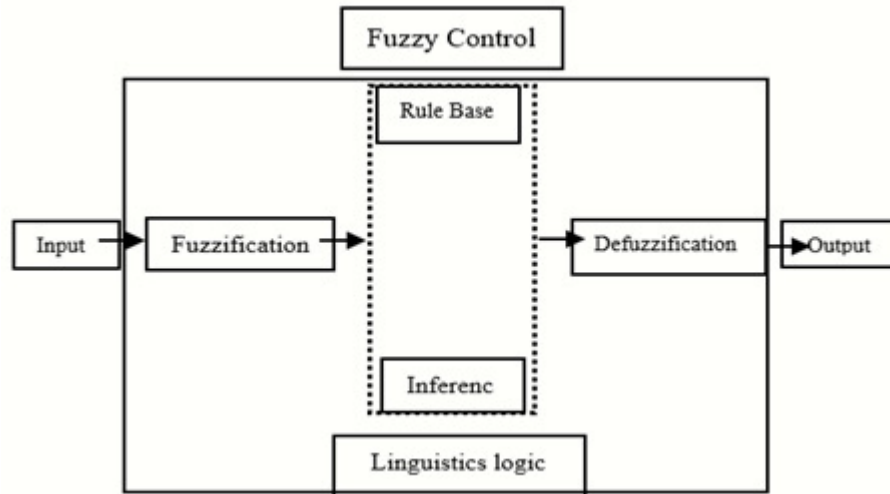


Figure 2: Fuzzy Logic General Architecture

### 5.2.1-Fuzzification

receives the inputs and transforms it into fuzzy sets (by maps the given inputs to fuzzy sets and linguistic variable). the fuzzy rules premises that are predefined in the rule base of the specified application can be matched with defined linguistic variables [36].

### 5.2.2- Rule Base

includes a set of linguistic rules designed in the form (if-then rules). These rules define the consequent of the model in terms of the given linguistic variables such as low, moderate and high. In addition, it specifies the type and number of the used membership functions of input and output parameters [38].

### 5.2.3-the fuzzy inference engine

is considered the central part of the fuzzy system it contains the Knowledge Base, that stores IF-THEN rules provided by experts. It simulates the human reasoning process by making fuzzy inference on the inputs and IF-THEN rules, combines rules from the fuzzy rule base and gives a mapping from input fuzzy sets, each rule is interpreted as a fuzzy implication. In this step the inference engine is applied to a set of rules included in the fuzzy rule base and that to produce the intended output. This procedure involves many steps as follows: first, it matches the linguistic variables of the input with the rules' premises. Second, it activates the matched rules in order to deduce the resultant of each fired rule, and finally it combines all consequents by using fuzzy set union in order to generate the final output which is represented as fuzzy set output [37].

#### 5.2.4- the Defuzzifier

It transforms the fuzzy set obtained by the inference engine into a crisp value, the output is produced as a linguistic variable, which is fuzzy and can be interpreted in different ways. The process of finding a crisp output after fuzzification and inference is called Defuzzification. Therefore, the fuzzy set output in this stage is converted to a crisp output (which results on a fuzzy output set).

### 5.3 Advantages of Using Fuzzy Logic with QA System

QA Question answering system enable users to access the knowledge in a natural way by asking questions and get back relevant correct answers, and that requires understanding of natural language text, linguistics and common knowledge, derived from the documents and providing accurate answers to a user's question often expressed as uncertainty words. So many of previous researchers on how to build QA logics, find that using artificial intelligence (AI) based methods that integrate natural language processing (NLP) techniques and knowledge base or corpus help to build QA logics. And they found that using fuzzy logic to modify linguistic values can retrieve the answers by matching user's question with the existing hierarchical ontology, and not only provide syntactic answers, but also semantic answers based on the question terms. It can produce result sets according to the degree of vague expressions in natural language, and vague adjectives in particular. And also, one of the main strengths of Fuzzy logic is that it allows the semantic partitions to overlap. In (1988) Smithson (argues that psychological explanations that permit choice under partial and uncertain constraints can be compatible with fuzzy logic [39]. For that Fuzzy logic is widely used in a different field such as expert systems, business and medicine. It can measure the output of the process and takes control actions on the process continuously.

Using fuzzy logic in NLP make the possibility of constructing a theory for artificial languages whose terms have fuzzy meaning and contribute it to a clarification of the concept of semantic meaning. Fuzzy logic, has a direct relation to NLP and QA development, A Critical Survey on the use of Fuzzy in Natural Language Processing start from (2004, Araujo) for rule learning, in 2010 was for knowledge representation by and in 2011 by Lai, Wu, Lin, & Huang.), research was for word meaning inference in (2012, Carvalho, Batista, & Coheur) for linguistic summarization[38]. in 2013 was two search the first one was about grammatical inference, by (Kazemzadeh, Lee, & Narayanan) and the second one was for emotion recognition by (Luong, Socher, & Manning) in 2014 by (Cambria, Gastaldo, Bisio, & Zunino) for text categorization. Most of the researchers currently working in the NLP field with fuzzy logic in different aspects, tries to find some guidelines on what could be done to and make fuzzy logic more interfere with NLP. Therefore, many of the previous researchers relied on used fuzzy logic in NLP have many advantage as it can represent the meaning of dispositions provide with more accuracy in expressive with easy-to-use operators provides simpler framework that lead to increase the accuracy of recognition system efficient. and also, fuzzy logic can use as translators or framework to match scores higher than before across different text domains.

## 6. CONCLUSIONS

In this survey paper are show an overview on what Question-Answering is and its architecture and the possible relationship with fuzzy logic and how fuzzy logic can be used as intermediate semantic representations of vague expressions. we pin down the previous related research with summarized and organized recent research results in a novel way that integrated and added

understanding to work in the question-answering field. It emphasized the classification of the existing literature, developing a perspective on the area, and evaluating trends. However, because it is impossible for a survey to include all or even most of previous research, this survey included only the work of the top-publishing and top-cited authors in the QA field.

## 7. FUTURE WORK

In this paper it has been discussed some of the elementary approach, and it perform fairly well but suffer from some limitations. This fact leads us to next step as a next step, we will try to tame our feature set on one possibility is to development a QA system using hybrid approach in our system that will be designed to be able to answer number of questions mainly based on fuzzy rule that is less impacted by the serious imbalance between negative and positive instances. This should help to increase the accuracy of Future result the architecture of Future Question Answering system using fuzzy logic rule is as shown in Figure 3.

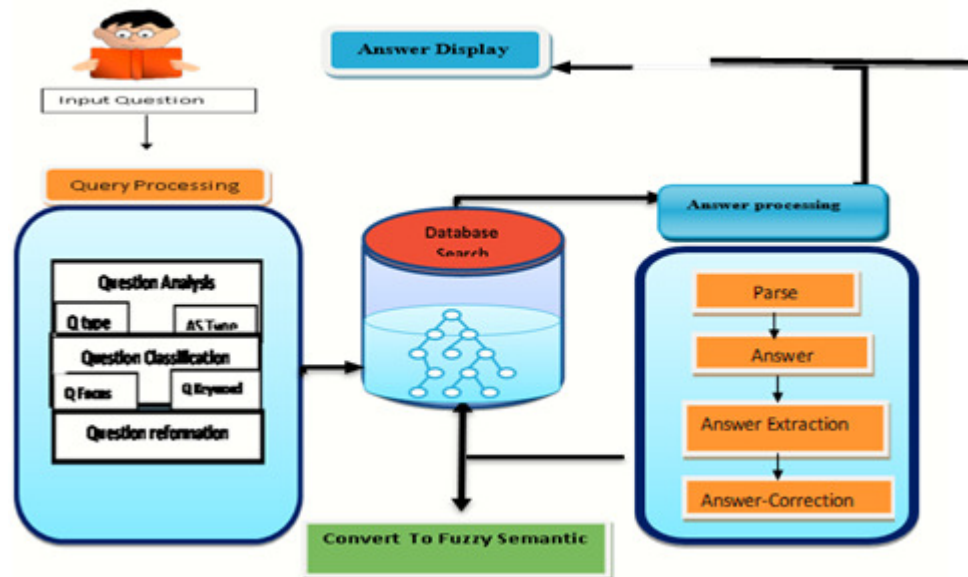


Figure 3. Question Answering System Using Fuzzy Logic Rule

## REFERENCES

- [1] Hendrix. G.G, Sacerdoti, E.D,sagalowicz. D. Slocum. J. "Developing a natural Language interface to complex data in ACM Transaction on database system. 3(2). pp. 105- 147,1978.
- [2] Green BF, Wolf AK, Chomsky C, and Laughery K. Baseball: An automatic question answerer. In Proceedings of Western computing Conference, Vol. 19, 1961, pp. 219–224.
- [3] [Cook, 2006] Cook, J. L. (2006). College students and algebra story problems: Strategies for identifying relevant information. Reading Psychology, 27:95 – 125.
- [4] Weizenbaum J. ELIZA - a computer program for the study of natural language communication between man and machine. In Communications of the ACM, Vol. 9(1), 1966, pp. 36-45

- [5] Woods W. Progress in Natural Language Understanding - An Application to Lunar Geology. In Proceedings of AFIPS Conference, Vol. 42, 1973, pp. 441–450.
- [6] Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. (1977). GUS, A frame driven dialog system. *Artificial Intelligence*, 8, 155–173.
- [7] Julian Kupiec. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In Robert Korfage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 181–190. Special issue of the SIGIR FORUM
- [8] Lehnert, W. G. (1978). *The Process of Question Answering: A Computer Simulation of Cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- [9] Huettner, A. (2000) “Question Answering”. In: 5th Search Engine Meeting.
- [10] Voorhees EM. The TREC-8 question answering track report. In *Proceedings of TREC-8*, 1999, pp. 77-82.
- [11] K.L.Kwok,L.Grunfeld,N.Dinstl,and M.Chan.TREC-9 Cross language,web and quationt -answering track experiments using PIRCS.In Voorhees and Harman .
- [12] E.M.Voorhees and D.K. Harman,editors,Proceedings of the Ninth Text REtrieval Conference (TREC-9)
- [13] Stefanie Tellex, "Pauchok: A Modular Framework for question Answering", Master Thesis Submitted to the Department of Electrical Engineering and computer science, Maccachusetts institute of Technology, June 2003..
- [14] Michael Kaisser. Question Answering by Searching Large Corpora with Linguistic Methods. Master's thesis, Saarland University, Germany, 2004.
- [15] Riloff E and Thelen M. A Rule-based Question Answering System for Reading Comprehension Tests. In *ANLP/NAACL Workshop on Reading Comprehension Tests as*.
- [16] Ittycheriah A, Franz M, Zhu WJ, Ratnaparkhi A and Mammone RJ. IBM's statistical question answering system. In *Proceedings of the Text Retrieval Conference TREC-9*, 2000.
- [17] Joho, H. (1999) "Automatic detection of descriptive phrases for Question Answering Systems: A simple pattern matching approach". MSc Dissertation. Department of Information Studies, University of Sheffield. Sheffield, UK
- [18] Deepak Ravichandran, Abraham Ittycheriah and salimroukos, "Automatic Derivation of surface text pattern for a maximum Entropy Based question answering system", Work done while the author was an intern at IBM TJ Watson research center during summer 2002.
- [19] Cai D, Dong Y, Lv D, Zhang G, Miao X. A Web-based Chinese question answering with answer validation. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp. 499-502, 2005.
- [20] J.-L. Binot, L. Debillé, D. Sedlock, and B. Vandecapelle. *Natural Language Interfaces: A New Philosophy*. *SunExpert Magazine*, pages 67{73, January 1991

- [21] Berger A, Caruana R, Cohn D, Freitag D, and Mittal V. Bridging the lexical chasm: statistical approaches to answer-finding. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 192-199.
- [22] Bronnenberg, W.J., Bunt, H.C., Landsbergen, S.P.J., Scha, RoJ.H., Schoenmakers, W.J., van Utter,n, E.P.C. (1979) The question answering system PHLIQAI. In L.Bolc (ed.), Natural communication with computers, McMillan, London; Hanser Verlag, M~nehen.
- [23] SACERDOTI, E.D. Language access to distributed data with error recovery. Proc. 5th Int. Joint Conf. on Artificial Intelligence, Cambridge, Mass., Aug. 1977.
- [24] Ask Jeeves. 1996. www.ask.com Site last visited in 28-march-2018.
- [25] AnswerBus, Question Answering System. Website: <http://answerbus.com> J. Allen. 1995. Natural Language Understanding. The Benjamin/Cummings Publishing Company, Menlo Park, CA.
- [26] Vargas-Vera M. and Motta E and Domingue J. (2003a): AQUA: An Ontology-Driven Question Answering System. AAAI Spring Symposium, New Directions in Question Answering, Stanford University, March 24-26, 2003.
- [27] Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimmy Lin, Gregory Marton, Alton Jerome McFarland, and BarisTemelkuran. 2015. Omnibase: Uniform access to heterogeneous data for question answering. In Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2015).
- [28] Unmeshsasikumar, Sindhu L, "A survey of Natural Language question answering system", international journal of computer applications(0975-8887), volume 108 -No 15. December 2014 .
- [29] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pages 1156–1165, New York, NY, USA. ACM
- [30] Sanjay K. Dwivedi and vaishalisingh "Research and reviews in question answering system", International Conference on Computational Intelligence: Modeling Techniques and Applications CIMTA) 2013 Procedia Technology 10 ( 2013 ) 417 – 424.
- [31] Ellen M. Voorhees and Dawn M. Tice. 1999. The TREC-8 question answering track evaluation. TREC-8 Proceedings, pages 41–63, Nov.
- [32] JIPING SUN, FAKHRI KARRAY, Fuzzy Logic-Based Natural Language Processing and Its Application to Speech Recognition, Ontario, N2L 3G1, Canada,2017
- [33] J.Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '13), vol. 2, p. 6, 2013. View
- [34] F. J. Estrella, M. Espinilla, F. Herrera, and L. Martínez, "FLINTSTONES: a fuzzy linguistic decision tools enhancement suite based on the 2-tuple linguistic model and extensions," Information Sciences, vol. 280, pp. 152–170, 2014
- [35] V. C. Gerogiannis, E. Rapti, A. Karageorgos, and P. Fitsilis, "A fuzzy linguistic approach for human resource evaluation and selection in software projects," in Proceedings of the 5th International Conference on Industrial Engineering and Operations Management (IEOM' 15), pp. 1–9, Dubai, UAE, March 2015.

- [36] VAN RIJSBERGEN, C.J. Information Retrieval, Second Edition, Butterworths, London (1979).
- [37] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages", Comput. Math. Appl., vol. 9, pp. 149-184, 1983.
- [38] A. Niewiadomski, "A type-2 fuzzy approach to linguistic summarization of data", IEEE Trans. Fuzzy Syst., vol. 16, no. 1, pp. 198-212, Feb. 2008.
- [39] Vilém Novák, Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on, 10.1109/FUZZ-IEEE.2017.8015405, Naples, Italy, 8, 2017
- [40] Zadeh, L. A. (1978) ,PRUF-A Meaning Representation Language for Natural Languages, Int. Journal Man–Machine Studies, 10, pp. 395–460; also in Fuzzy Sets and Applications: Selected Papers by L. A. Zadeh, John Wiley & Sons, New York, pp. 499–568 (1987).
- [41] Konstantinos Chatzitheodorou, Proceedings of the Workshop on Natural Language Processing (NLP4TM) Improving translation memory fuzzy matching by paraphrasing, pages 24–30, Hissar, Bulgaria, Sept 2015
- [42] L. Zadeh. Test-Score Semantics as a Basis for a Computational Approach to the Representation of Meaning. Literary and Linguistic Computing, 1986.
- [43] Fenza, G., V. Loia, and S. Senatore. 2008. "Concept Mining of Semantic Web Services by Means of Extended Fuzzy Formal Concept Analysis (FFCA)." IEEE Int. Conf. on Systems, Man and Cybernetics
- [44] Monika Rani, Maybin K. Mueyba, O.P. Vyas, A hybrid approach using ontology similarity and fuzzy logic for semantic question answering." In Advanced Computing, Networking and Informatics Volume 1, pp. 601-609. Springer, Cham, 2014.

## AUTHOR

Eman Mohamed Nabil Alkholy

Faculty of Computers & Information, Helwan University, Cairo, Egypt.



*INTENTIONAL BLANK*



# PROMOTING STUDENT ENGAGEMENT USING SOCIAL MEDIA TECHNOLOGIES

Mohammad Alshayeb

Information and Computer Science Department  
King Fahd University of Petroleum & Minerals  
Dhahran 31261, Saudi Arabia

## **ABSTRACT**

*Using social media in education provides learners with an informal way for communication. Informal communication tends to remove barriers and hence promotes student engagement. This paper presents our experience in using three different social media technologies in teaching software project management course. We conducted different surveys at the end of every semester to evaluate students' satisfaction and engagement. Results show that using social media enhances students' engagement and satisfaction. However, familiarity with the tool is an important factor for student satisfaction.*

## **KEYWORDS**

*Student engagement, social media, instructional technology*

## **1. INTRODUCTION**

Instructors in higher education have been using different computer related technologies in education, such as multimedia [1, 2], communication [3] and mobile [4] to enhance students' learning. Different technologies have been utilized in the classroom environment such as blogs, wikis, portals, instant messaging, and Facebook [5-7].

Researchers reported that the use of technology and electronic media has improved teaching and learning process [8-11]. Babur [12] found that using technology in teaching helped students to increase their capability for understanding. Beichner et al. [13] also found that students taught with a technology-based approach outperformed and were more satisfied than students taught with the traditional teaching methods. Mobile technologies were also found to be effective in teaching and learning [14, 15]. Barbosa et al. reported that the use of mobile in teaching improved the learning and the interaction between learners [16]. Game-based learning [17, 18] was also found to be effective in enhancing students' learning process. However, other researchers found that technology may have negative effects [19-24].

Student engagement is important for students' performance, retention and achievement [25]. Manwaring et al. [26] investigated student engagement in blended learning classes. Gunuc and Kuzu [25] conducted a study to evaluate the impact of technology on student engagement and to

find the relationships between student engagement and technology use in class. They found that the use of technology in and out the class increased student engagement. Bray and Tangney [4] explored the impact of using mobile technologies in students' engagement; they found that the use of mobile technologies increases student engagement.

Social media tools provide interactive environment for communication. Computer communication methods not only should be usable, but they should also be engaging [27]. Dyson et al. [28] conducted an experiment with a large-scale class in which two-thirds of the class were supposed to use Facebook and one-third without. They found that students who did not view the Facebook postings reported lower engagement and understanding of the in-class discussion. Rashid and Asghar [29] conducted a study to evaluate the impact of technology use in student engagement. They found that the use of technology has a direct positive relationship with students' engagement. Imlawi et al. [30] reported that students are more engaged, motivated and satisfied when the instructor uses course-based online social networks to communicate with them.

In this paper, the researcher evaluates students' engagement when using three different social media technologies. The researcher first started by using Facebook for four years; during these four years, students reported that they were motivated and engaged, however, students after the four years, started to report that Facebook is not appropriate for communication and hence engagement has been reduced. Based on their recommendation and the instructor's evaluation, WhatsApp is used to replace Facebook. WhatsApp was used successfully for one year were students were positive about its use, however, due to the need for personalized communication, Slack [31] tool was selected. Results are promising, even though students were more satisfied in WhatsApp than Slack, yet they have shown interest and satisfaction.

The reminder of this paper is organized as follows: Section 2 describes the course and the communication objectives. Section 3 presents the rational for selecting the used tools. The evaluation and discussion of student engagement and satisfaction is presented in section 4. Finally, section 5 presents the conclusion and the future work.

## **2. COURSE DESCRIPTION AND COMMUNICATION OBJECTIVES**

In this section, we discuss the course description and the objectives of using different social media technologies to promote students' engagement.

### **2.1 Course Description**

Using technology in a course should serve the course objective. An important course objective in the software project management course is enhancing communication. The availability of software technology tools can help in achieving this objective. Software project management is a core course in the software engineering program at King Fahd University of Petroleum and Minerals (KFUPM) and can be taken as an elective for other students in other disciplines. Students are required to do a project and three homework assignments that require teamwork and hence good communication skills is required. In addition, the project and the assignments require lots of interaction with the course instructor. All students registered in the course are from IT background (software engineering, computer science and computer engineering) and hence it is easy to use software technologies.

## 2.2 Objectives of using Communication Tools

As indicated earlier, the course requires lots of communication among the students themselves and between the instructor and the students. The following are the different communication objectives have been achieved by using social media technologies:

- **Course Announcements:** Posting course announcement for quick dissemination. This includes all announcements related to the course such as homework deadlines and change in the plan etc.
- **Discussion Questions by the Instructor:** Posting questions for discussion for students. Since most students are often online, many of them participated effectively.
- **Discussion Questions by the Students:** To motivate effective problem solving and team work, each team was assigned for one week the responsibility for answering all the questions posted. This new twist encouraged students to ask/answer the discussion questions.
- **Chat/Virtual Office Hours:** Before the exam, online office hours session was conducted so that students can chat with the instructor to answer their questions.
- **Response to Urgent Queries:** since the social media are accessible anywhere any time, this enabled quick response to all students' queries no delay especially before the submission deadlines.

## 3. SOCIAL TOOL SELECTION

The use of social media technologies in my classes was started in 2011 by using Facebook. The selection of Facebook was based on my evaluation of the appropriate tool in addition to students' request. At the end of every semester, the course instructor conduct a survey to evaluate students' engagement and satisfaction of using the technology. In 2014 survey, most students indicated that the Facebook is not appropriate to be used anymore, specific students; comments are listed below:

- Unfortunately, Facebook is dying in our region and I only open it for this group
- I do not recommend Facebook in the next offering
- To me, the use of Facebook was really a bad experience
- It is better to use technology other than Facebook
- NOT using Facebook
- It was hard for me to check on Facebook, also using WhatsApp group could be a good and faster way

After this feedback and based on my evaluation for appropriate tools, the course instructor stopped using Facebook and started using WhatsApp in 2014. Even though WhatsApp use was successful, there was a need for personalized team communication which is not supported by WhatsApp. Hence, it was decided to search for a more appropriate technology and started to use Slack [31] as a new method for communication.

## 4. EVALUATION

As indicated in section 3, at the end of each semester, the course instructor conduct a survey to evaluate student's engagement and satisfaction. The details of surveys to evaluate these social media technologies are discussed in the following subsections.

### 4.1 Facebook

Facebook was first used in my classes in 2011. The details of the survey that evaluate the use of Facebook are discussed in this section. In response to question "The use of Facebook for class communication was effective and useful and enhanced my learning experience", 95% of the students either strongly agree or agree that using Facebook for class communication was effective and useful and enhanced their learning experience as shown in Figure 1. In response to a question "Which technologies, used in the course, improved your learning the most?", 15 students out of 22 selected Facebook.

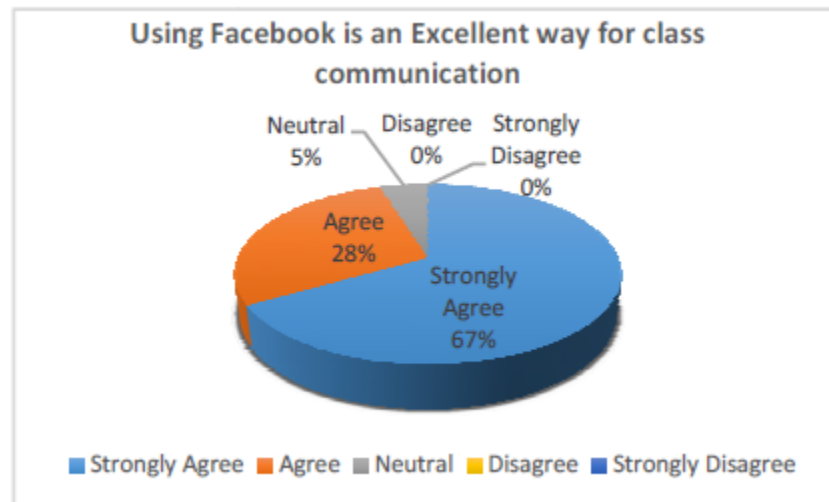


Figure 1 : Using Facebook for communication

Figure 3 shows that 83% of the students either strongly agree or agree that discussion questions posted on that WhatsApp group were useful.

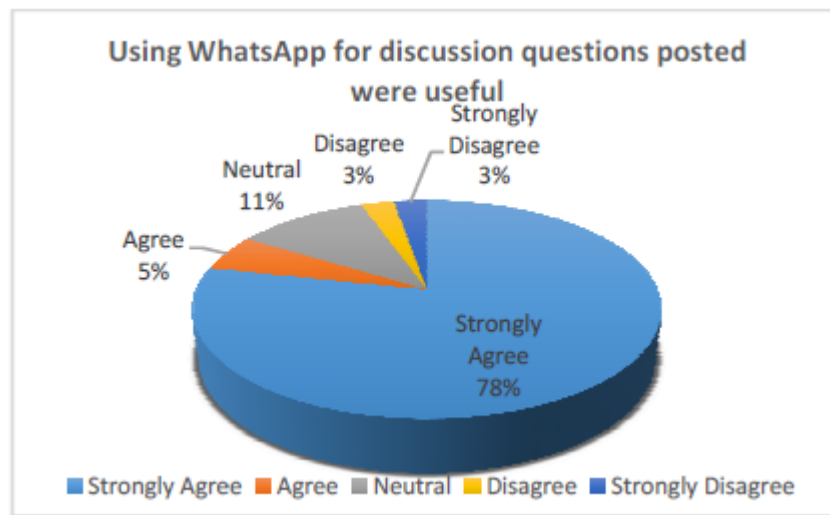


Figure 3. Using WhatsApp for discussion questions

Figure 4 shows that 87% of the students recommend using WhatsApp in the future offering of the course.

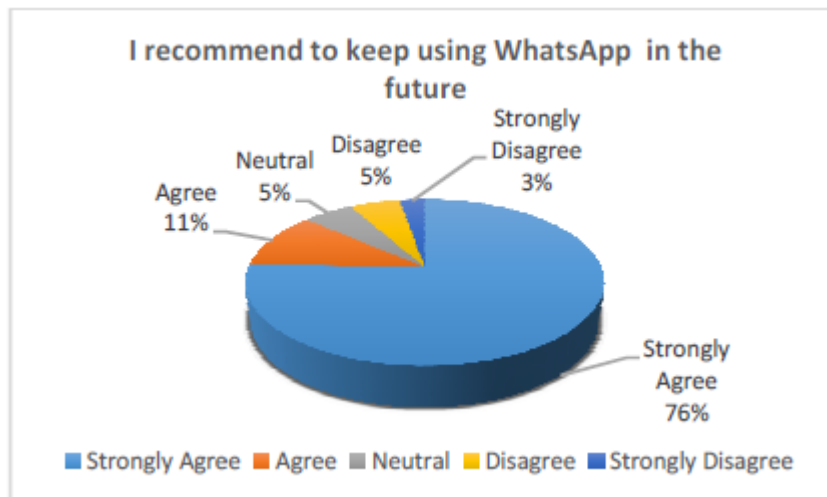


Figure 4. Recommending using WhatsApp in the Future

Even though students were engaged and satisfied by using WhatsApp, there was a need personalized team communication; to do that in WhatsApp, a group for each team should be created which is not convenient. Hence, the instructor searched for an alternative tool to be used and selected Slack.

### 4.3 Slack

Slack is a tool that is mainly used to manage projects. For communication, it has all features of WhatsApp, in addition it can have a personalized channel for teams' communication. Furthermore, a wide number of tools can be integrated with Slack. Therefore, slack was selected

mainly for the personalized communication. Slack was used so far for one semester; a survey was conducted in 2018. Figure 5 shows the student response for using slack as a communication method. 65% of the students agree or strongly agree that using slack is an excellent way of communication. This percentage is less than for WhatsApp. When investigating the reasons for that, students reported that this is the first time of using slack, therefore they are not familiar with it.

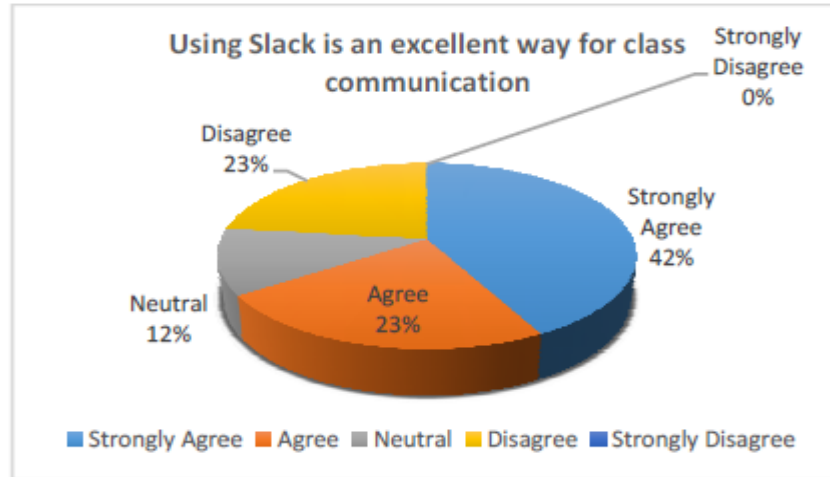


Figure 5. Communication Using Slack

#### 4.4 Technology Use

In addition to the surveys for the social media technologies, questions on using technology in general are asked to the students. In 2011, in response to the question "Using technology in teaching this course enhanced my learning experience". Figure 6 shows that 100% of the students either strongly agree or agree that using technology in teaching this course enhanced their learning experience.

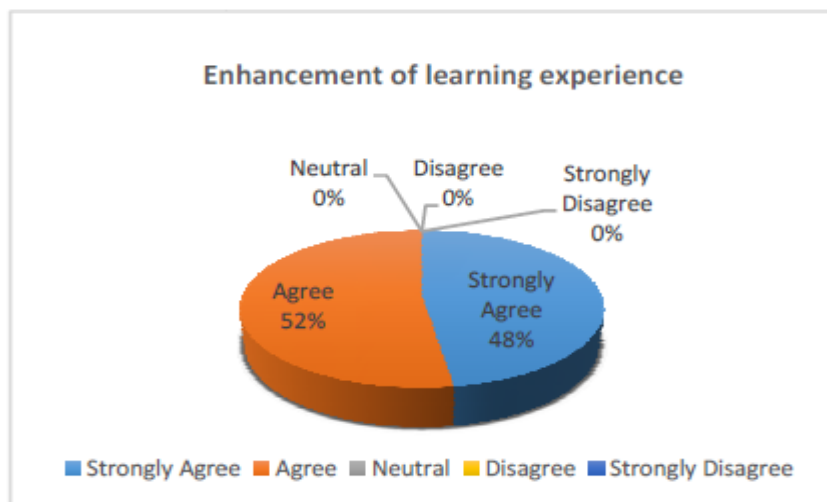


Figure 6: Using technology in enhancing the learning experience

In 2015, the survey included more questions to evaluate students' engagement and satisfaction. In response to the question: "The use of technology made the course more interesting", 100% of the students either strongly agree or agree that using technology in the course made the course more interesting as shown in Figure 7.

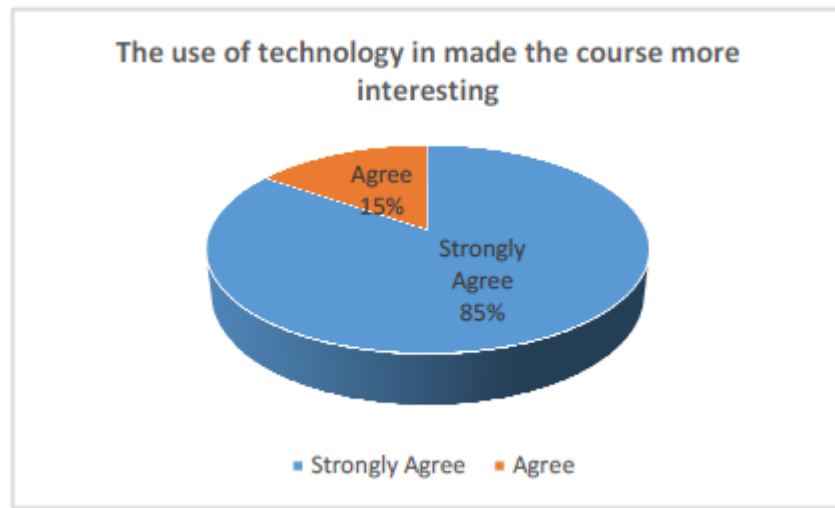


Figure 7: The use of technology in made the course more interesting

In response to the question: "The use of technology made me more interested in the course", 100% of the students either strongly agree or agree that using technology in the course made them more interested in the course as shown in Figure 8.

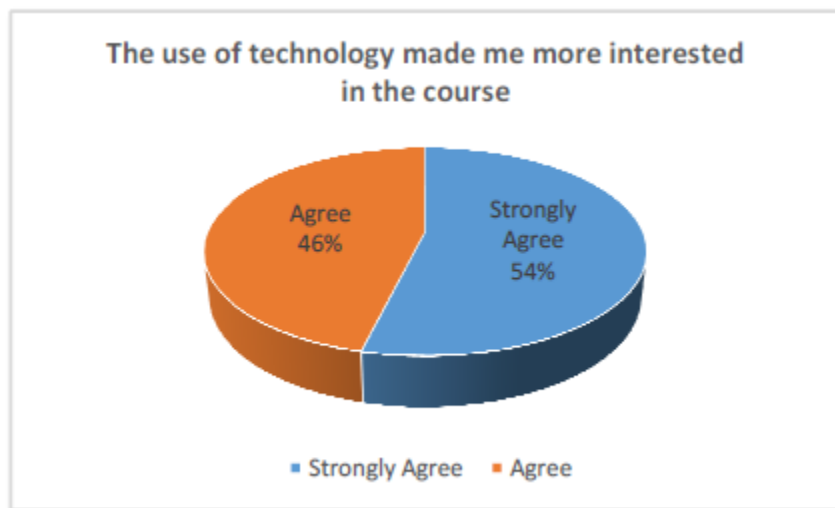


Figure 8: The use of technology and interest in the course

92% of the students either strongly agree or agree that using technology in the course enabled effective communications among students as shown in Figure 9.

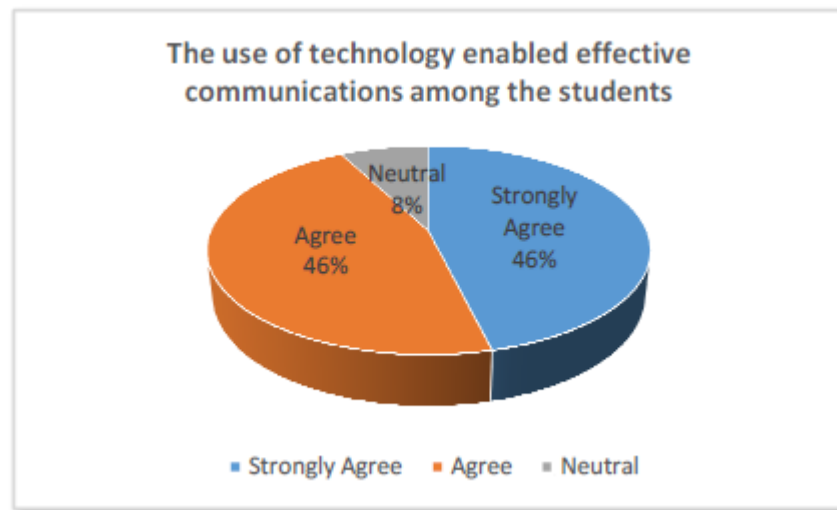


Figure 9: The use of technology and effective communications among students

100% of the students either strongly agree or agree that using technology in the course enabled effective communication between the students and the instructor as shown in Figure 10.

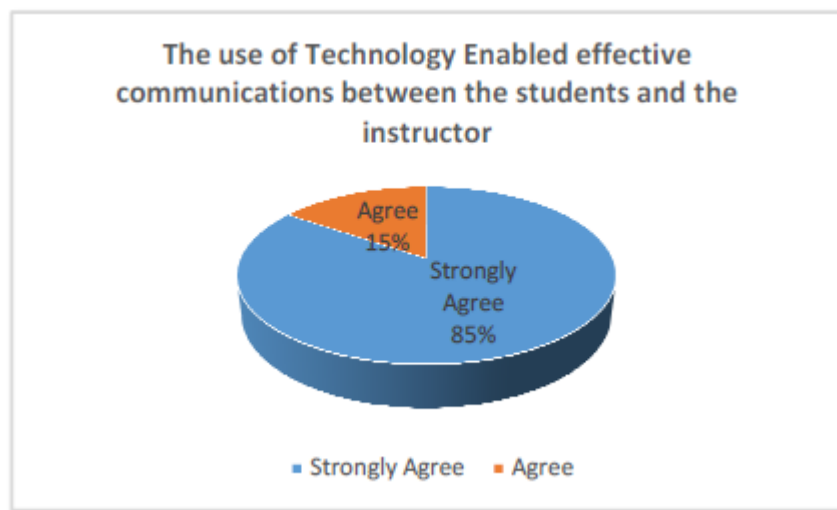


Figure 10: The use of technology enabled effective communications between the students and the instructor

100% of the students strongly agree recommend using different technologies in the future offering of the course as shown in Figure 11.



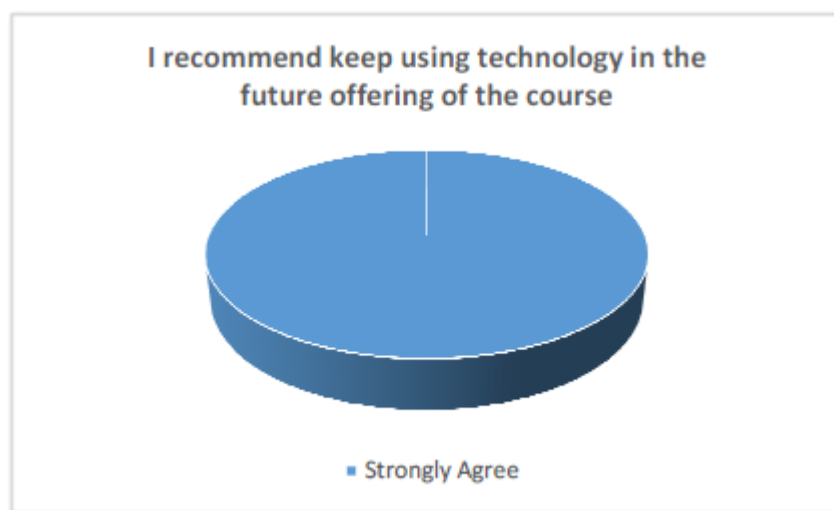


Figure 11: Recommendation of using technology in the future

#### 4.5 Benefits of using Social Media Technologies

Using the above mentioned social media technologies provided a better students' learning experience and promoted students' engagement. This can be attributed to many reasons as these tools : 1) enhanced the communication between the instructor and students and among students, 2) extended the communication for more than the lecture time, 3) provided the students with a convenient way to ask and answer discussion questions, finally, 4) created a referable record of course discussions. In addition, they helped the instructor to identify concepts that need further discussion and helped in enhancing critical thinking by students.

### 5. CONCLUSION

Social media tools provide interactive environment for communication. This paper presents our experience in using three different social media technologies to enhance students' engagement. We presented our experience with using Facebook for four years then moved to use WhatsApp based on students' request and the instructor's evaluation. Finally, Slack tool was used as a new method for communication.

Our results show that the three methods were found to be effective in promoting students' engagement. However, students seem to be less satisfied when using Slack. This is attributed to the fact that most students never used it before and hence they were not familiar with it, yet, majority were satisfied.

In our future studies, we plan to evaluate the impact of students' performance in using social media tools and technologies, in addition, we plan to evaluate the impact of using gamification technologies in students' engagement and performance.

### ACKNOWLEDGEMENT

The authors acknowledge the support of King Fahd University of Petroleum and Minerals.

**REFERENCES**

- [1] E. Cutrim Schmid, "Potential pedagogical benefits and drawbacks of multimedia use in the English language classroom equipped with interactive whiteboard technology," *Computers & Education*, vol. 51, no. 4, pp. 1553-1568, 2008
- [2] L. Proserpio and D. Gioia, "Teaching the virtual generation," *Academy of Management Learning and Education*, vol. 6, no. 1, pp. 69-80, 2007.
- [3] M. Alavi and R. B. Gallupe, "Using Information Technology in Learning: Case Studies in Business and Management Education Programs," *Academy of Management Learning & Education*, vol. 2, no. 2, pp. 139-153, 2003.
- [4] A. Bray and B. Tangney, "Enhancing student engagement through the affordances of mobile technology: a 21st century learning perspective on Realistic Mathematics Education," *Mathematics Education Research Journal*, journal article vol. 28, no. 1, pp. 173-197, March 01, 2016
- [5] S. L. Connell, "Comparing blogs, wikis, and discussion boards as collaborative learning tools," in "In Wiki, Hyderabad: India: ICFAI (the Institute of Financial Analysts of India) University Press, "2006.
- [6] J. Farmer, "Communication dynamics: Discussion boards, weblogs and the development of communities of inquiry in online learning environments," in *Conference of the Australasian Society for Computers in Learning in Tertiary Education*, 2004, pp. 1-10.
- [7] D. Fichter, "The many forms of e-collaboration: Blogs, wikis, portals, groupware, discussion boards, and instant messaging," *Trade Publication*, vol. 29, no. 4, pp. 48-50, 2005.
- [8] E.-S. Aziz, "Teaching and Learning Enhancement in Undergraduate Machine Dynamics," *Computer Applications in Engineering Education*, vol. 19, no. 2, pp. 244-255, 2011.
- [9] E. Y. Borkowski, D. Henry, L. L. Larsen, and D. Matelik, "Supporting teaching and learning via web: Transforming hard-copy linear mind sets into web flexible creative thinking," *Journal of Network Computer Application*, vol. 20, pp. 253-265, 1997.
- [10] A. F. Grasha and H. N. Yangarber, "Integrating teaching styles and learning styles with instructional technology," *College Teaching*, vol. 48, no. 1, pp. 2-9, 2009.
- [11] A. K. Aggarwal and R. Bent, "Web based education, In learning and teaching Technologies," *Web-Based Opportunities and Challenges*, pp. 2-16, 2000.
- [12] B. Deliktas, "Computer technology for enhancing teaching and learning modules of engineering mechanics," *Computer Applications in Engineering Education*, vol. 19, no. 3, pp. 421-432, 2011.
- [13] R. Beichner et al., "Case study of the physics components of an integrated curriculum. ," *American Journal of Physics*, vol. 67, pp. S16-S24, 1999.
- [14] C. Romero, S. Ventura, and P. d. Bra, "Using mobile and web-based computerized tests to evaluate university students," *Computer Applications in Engineering Education*, vol. 17, no. 4, pp. 435-447, 2009.
- [15] G. Vavoula and M. Sharples, "Challenges in evaluating mobile learning," in *Proceedings of the mLearn 2008 Conference*, Shropshire, United Kingdom, 2008.

- [16] J. L. V. Barbosa, R. Hahn, D. N. F. Barbosa, and W. Segatto, "Intensive use of mobile technologies in a computer engineering course," *Computer Applications in Engineering Education*, 2012.
- [17] W. R. Watson, C. J. Mong, and C. A. Harris, "A case study of the in-class use of a video game for teaching high school history," *Computers & Education*, vol. 56, no. 2, pp. 466-474, 2011.
- [18] M. Ebner and A. Holzinger, "Successful implementation of user-centered game based learning in higher education: An example from civil engineering," *Computers & Education*, vol. 49, no. 3, pp. 873-890, 11// 2007.
- [19] V. K. G. Lim, "The IT way of loafing on the job: Cyberloafing, neutralizing and organizational justice," *Journal of Organizational Behavior*, vol. 23, no. 5, pp. 675-694, 2002.
- [20] M. J. Austin and L. D. Brown, "Internet plagiarism: Developing strategies to curb student academic dishonesty," *The Internet and Higher Education*, vol. 2, no. 1, pp. 21-33, 1999.
- [21] D. L. McCabe, K. D. Butterfield, and L. K. Treviño, "Academic dishonesty in graduate business programs: Prevalence, causes, and proposed action," *Academy of Management Learning and Education*, vol. 5, no. 3, pp. 294-305, 2006.
- [22] R. E. Mayer, J. Heiser, and S. Lonn, "Cognitive constraints on multimedia learning: When presenting more materials results in less understanding," *Journal of Educational Psychology*, vol. 93, no. 1, pp. 187-198, 2001.
- [23] S. C. Rockwell and L. A. Singleton, "The effect of the modality of presentation of streaming multimedia on information acquisition," *Media Psychology*, vol. 9, no. 1, pp. 179-191, 2007.
- [24] C. Chou, "Internet Heavy Use and Addiction among Taiwanese College Students: An Online Interview Study," *CyberPsychology & Behavior*, vol. 4, no. 5, pp. 573-585, 2004.
- [25] S. Gunuc and A. Kuzu, "Confirmation of Campus-Class-Technology Model in student engagement: A path analysis," *Computers in Human Behavior*, vol. 48, pp. 114-125, 2015/07/01/2015.
- [26] K. C. Manwaring, R. Larsen, C. R. Graham, C. R. Henrie, and L. R. Halverson, "Investigating student engagement in blended learning settings using experience sampling and structural equation modeling," *The Internet and Higher Education*, vol. 35, pp. 21-33, 2017/10/01/ 2017.
- [27] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the Association for Information Science and Technology*, vol. 61, no. 1, pp. 50-69, 2010.
- [28] B. Dyson, K. Vickers, J. Turtle, S. Cowan, and A. Tassone, "Evaluating the use of Facebook to increase student engagement and understanding in lecture-based classes," *Higher Education*, journal article vol. 69, no. 2, pp. 303-313, February 01 2015.
- [29] T. Rashid and H. M. Asghar, "Technology use, self-directed learning, student engagement and academic performance: Examining the interrelations," *Computers in Human Behavior*, vol. 63, pp. 604-612, 2016/10/01/ 2016.
- [30] J. Imlawi, D. Gregg, and J. Karimi, "Student engagement in course-based social networks: The impact of instructor credibility and use of communication," *Computers & Education*, vol. 88, pp. 84-96, 2015/10/01/ 2015.
- [31] Slack. Available: <https://slack.com/>

*INTENTIONAL BLANK*

# MOVING FROM WATERFALL TO AGILE PROCESS IN SOFTWARE ENGINEERING CAPSTONE PROJECTS

Mohammad Alshayeb<sup>1</sup>, Sajjad Mahmood and Khalid Aljasser

King Fahd University of Petroleum and Minerals  
Dhahran 31261, Saudi Arabia.

## ABSTRACT

*Universities offer software engineering capstone course to simulate a real world-working environment in which students can work in a team for a fixed period to deliver a quality product. The objective of the paper is to report on our experience in moving from Waterfall process to Agile process in conducting the software engineering capstone project. We present the capstone course designs for both Waterfall driven and Agile driven methodologies that highlight the structure, deliverables and assessment plans. To evaluate the improvement, we conducted a survey for two different sections taught by two different instructors to evaluate students' experience in moving from traditional Waterfall model to Agile like process. Twenty-eight students filled the survey. The survey consisted of eight multiple-choice questions and an open-ended question to collect feedback from students. The survey results show that students were able to attain hands on experience, which simulate a real world-working environment. The results also show that the Agile approach helped students to have overall better design and avoid mistakes they have made in the initial design completed in of the first phase of the capstone project. In addition, they were able to decide on their team capabilities, training needs and thus learn the required technologies earlier which is reflected on the final product quality.*

## KEYWORDS

*Agile Process, Waterfall Process, Capstone Project, Software Engineering*

## 1. INTRODUCTION

IEEE computer society and the Association for Computing Machinery (ACM) recommend that software engineering students undertake a capstone project that integrates knowledge of software development life cycle, learned throughout the course of the undergraduate software engineering program, in a realistic simulation of professional experience. The traditional waterfall driven approach advocates that a well-planned process for capstone projects would be efficient approach for inexperienced software engineering students. On the other hand, agile driven approach provides hands-on environment to learn and apply software development life cycle knowledge and skills.

The main criteria for selecting one of these approaches is its effectiveness to help students in delivering a successful product at the end of the capstone project. We have been using Waterfall

---

<sup>1</sup>Corresponding author

model for twelve years. Waterfall model is known for its organized set of steps starting for the early stages of project proposal till the last stages of testing and delivery. On a 2-semester software engineering project (15 weeks in each semester), the duration is more than enough to cover all the stages of a software project. However, because of the structure of the course (shown in Section 2) and the required deliverables, students find themselves delayed and less motivated because they do not see their software products until the second half of the second semester. Waterfall model would be suitable if the software idea and its requirements are known upfront. In our capstone project, however, when following the Waterfall approach students face several challenges and difficulties. For example, with student little experience in design, after the implementation they realize that the code is not consistent with the design. Since they have only one cycle, it will be hard to go back and fix the design issues. This is even harder for the course instructor since the number of revisions on the design are different from one project to another. Furthermore, most students use new and emerging technologies which have not been discussed in the course they have already taken, therefore, when the implementation time comes, they have to learn new development frameworks and tools which hinder their speed of implementing the planned features which will delay the overall project. This will lead to little time for testing which can affect the final product quality. Finally, students indicated their dissatisfaction from the process as they only do documentation of the Software Requirement Specification (SRS) and the Software Design Document (SDD) in the first semester which causes lack of motivation towards remaining project tasks. This motivated us to move to use Agile in the implementation of the capstone project especially with the positive and promising results reported in literature for moving from Waterfall to Agile in capstone projects.

Over the last few years, a number of studies reported experience about the use of both traditional waterfall plan driven and agile driven approaches for software engineering capstone course [1-5]. Rover et al. [6] presented a case study of developing software applications in two-semester senior project. They found that the use of Agile was successful and beneficial. The results of the case study show that it leads to high satisfaction and helped in producing high quality software. Kuhl [7] reported his experience in moving from the traditional Waterfall model to Agile process. The approach is meant to replace the sequential, and documentation-intensive, steps of the waterfall model with shorter development cycles by releasing a small set of features in each cycle. The author reported that the average quality of the projects was improved when using Agile relative to using Waterfall method. This was obvious as teams were able to implement more functionality and a better level of testing. Ding et al. [8] presented a case study of using Agile and Scrum methods in capstone projects. He presented recommendations to consider when applying Agile in capstone projects. Coupal and Boechler[9] reported their experience and observation in using Agile process in capstone project. They observe that using Agile supports learning and provides a good learning experience and produces good quality products within an academic environment. Devedzic and Milenkovic[10] discussed how to overcome potential problems in teaching Agile along with some recommendation based on their experience in teaching Agile in different universities and different cultural settings. Mahnic[1] reported his experience in teaching undergraduate capstone course in software engineering using Agile. He provided an empirical evaluation of students' progress in estimation and planning. He also provided recommendations and lessons learned to consider when applying Agile in courses. The survey he conducted shows that the students were satisfied and the course met the expectations. These positive and promising results motivated us to conduct our own study in our own settings. Rico and Sayani[11] described their use of Agile methods in software engineering capstone course. With little training in agile methods, the three teams were able to complete fully functional e-commerce websites. Knudson and Radermacher[12] provided suggestions from interviews of developers and managers who use Agile process and from student feedback for incorporating agile processes in the capstone course. Souza et al. [13] presented an evaluation of adapting Scrum method to evaluate the capstone project. El-Abbassy et al. [14] presented a framework for adopting and evaluating agile practices in computer science education.

In this paper, we report on our experience and the changes made to move from Waterfall process to Agile process in conducting the software engineering capstone project. The modifications involve the restructuring of the course, the deliverables and the assessment plans. The paper also presents the survey results which indicate the strengths of using Agile process. Finally, discussion and some direction for further improvement in future offering of the course is discussed.

The paper is structured as follows. Section 2 discusses the course description and structure, Section 3 presents the Agile structure. Section 4 details the evaluation of the Agile adoption. Finally, section 5 presents the conclusion and future work.

## 2. COURSE DESCRIPTION AND STRUCTURE

The two-semester capstone project is designed to follow Waterfall development methodology in a real-world context. Students are required to work in teams of five to six members to develop a realistic project based on user requirements provided by industrial sponsors or domain experts. The course is allocated seven credit hours that are distributed over two semesters (each semester consists of fifteen weeks). Senior students who have completed a set of software engineering and computer science core courses including software requirements engineering, software design and project management register the capstone project. The capstone project has no formal lectures and overall course delivery is structured as a set of weekly deliverables. The project teams meet with the instructor every week to submit their deliverables. The two-semester capstone project has the following objectives:

1. To employ knowledge gained from courses throughout the program such as development of requirements, design, implementation, project management, and quality assurance to develop a software solution to a real-world problem.
2. To apply all appropriate project management techniques.
3. To learn how to work in teams.
4. To enhance communication and writing skills.
5. To instill life-long learning skills.
6. Understand the impact of computing solutions in a global and societal context.

The course delivery schedule and assessments are shown in Table 1.

Table 1: Course Schedule and Assessment

Semester	Deliverables	Assessment
First semester	Project proposal	5%
	Project Plan	10%
	Software requirements specification (SRS)	45%
	Software Design Document (SDD)	30%
	Test cases	5%
	Prototype	5%
Second Semester	SRS and SDD review	10%
	Project Plan	10%
	Implementation	50%
	Testing	10%
	User and Deployment Manuals	10%
	Release - Final delivery of system	10%

In the first semester, student work on the requirements and design of the project, they also develop the test cases. At the end, they develop a prototype. In the second semester, they review their requirement and design document and then they implement the system. After implementation, they test and write documentation. Finally, they release the product.

### 3. AGILE BASED PROJECT SETTING

The students enrolled in the waterfall-based project setting course were facing a number of challenges, as discussed in the introduction. The challenges were associated with student outcomes associated with design, implementation and selection of appropriate technology/tools for the projects. The main reason for these challenges were lack of experience with latest frameworks, tools and associated technologies. As a result, design specifications completed in the first semester of the project often required significant refactoring. Furthermore, teams need time to self-learn new technologies and platforms. Hence, teams were left with little time to work on the implementation and produce a system which satisfied the stakeholder requirements.

Table 2 shows the new agile based project setting. In the first part, students develop project plan, software requirements specification; software design document for 30% of features; implementation; testing and first release of the product. In the second part, students complete the design and implementation of the product, alpha and beta testing of their code and evaluate the final product.

Table 2: Agile based project schedule

First Semester															
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Project Plan		Requirements Specifications					Design of 30% features			Implementation				Testing & Release 1	
Second Semester															
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Requirements review	Design Specification					Implementation					Beta Testing		User Manual	Release 2	

After running the capstone project using Agile for two semesters, it was obvious that students have improved their skills in:

1. Self-learning: students were able to self-learn new development languages, tools and technologies early as they have to start implementation in the first semester. This allows students to have better understanding of the system requirements.
2. Better design: Students iteratively design the project and hence they were able to learn from the first cycle and avoid the mistakes in the second cycle.



3. Risk management: have a two released, students reduced the project risk as they have a working system early. As a result, teams have more confidence in their ability to meet stakeholder requirements and go through a through testing process before releasing the final product.
4. Project management: students have real work experience in Agile management and planning. Students were able to assess their capabilities early and hence prepare themselves for any needed training

#### 4. AGILE APPROACH EVALUATION

To evaluate the impact of using Agile process instead of traditional Waterfall model, a survey over two semesters for two different sections was conducted. In the first offering, 16 students participated in the survey while there are 13 students participated in the survey. The two offerings were done by two different instructors; however, they followed the same approach.

The survey consisted of 9 questions, the questions are listed below:

1. The experience in the first semester helped the team in enhancing the project design of the second phase.
2. Having design and implementation in the first semester, forced the team to learn the new technology I need earlier.
3. The experience gained in the first semester helped the team in better understanding the requirements of the project.
4. The experience gained in the first semester helped the team in better understanding the team capabilities.
5. The experience gained in the first semester helped the team in better understanding the needed training for the team.
6. The experience gained in the first semester helped our team in building a realistic schedule for the activities in the second semester.
7. The new setting helped the team in reducing the risk of not completing the project.
8. I would you recommend using the Agile approach in the future offering of the course?
9. Recommendations and Comments.

Figure 1 represents a summary of the student responses, the figure shows the average satisfaction for each survey question. Figure 1 clearly shown that all question scored at least 4.4 out of 5.



Figure 1. Satisfaction Survey Results

Below, we discuss the reasons for this high satisfaction.

**Question 1:** This is the first projects in which students perform the complete software life cycle. Students designed a system in their design course without full implementation, hence, they could not tell what works and what does not work in design. In this project, after developing the first phase, students are more confidence of how the design can translate to code and hence the design of the first part helped them in avoiding mistakes they have made in the design of the first part.

**Question 2:** Most students develop mobile applications for their senior project. As of now, the program does not offer such course. Hence, students have to learn different mobile development technologies in the first semester in order to implement the first cycle. This has a positive impact as they have a better learning curve when they reach the second semester to implement the remaining functionalities.

**Question 3:** With respect to requirements, students tends to understand the requirements more as they have to select 30% of them to implement, they usually select the most important requirements and thus analyze the requirements and have a better understanding to the who system requirements.

**Question 4 & 5:** Students form their own teams; therefore, they usually select other students whom they know, yet, students have different capabilities. When students are forced to implement part of the system, they have a better understanding of their capabilities and thus identify the needed training early. Identifying the needed training helped students to plan and get the required training earlier which speeded up their readiness for the next semester activities.

**Question 6:** Student lack planning experience, thus, their plan tends to optimistic in the first semester especially in the development part. In the second semester, they should have gained some experience to have more realistic schedule.

**Question 7:** Students have high satisfaction with regard to reducing the project since they have a working version form the first semester. This gives them more confidence and less pressure to complete the second part.

**Question 8:** High percentage of students recommend keeping using Agile in the future offering of the course. This overall result is obvious from the satisfaction of all previous questions.

### **Question 9: Recommendations and Comments**

Below are some recommendations and comments mentioned by the students:

- “I can't think of any, but this approach helped completing the project smoothly.”
- “Including one more agile cycles in 418, will be even better. Like dividing the remaining 70% to 30% - 40%.”
- “I liked the way of starting to implement as early as possible.”
- “It was great to develop the idea of the project in earlier stages.”
- “In my opinion, agile use was a success.”
- “I didn't try the old approach, but I imagine that this approach is way better, for all the points mentioned in the questions above. Many thanks to you and for the department's faculty for the continuous improvements you make.”
- “I think it would've been better if testing was required during the implementation phase, especially unit testing.”

It is obvious that students were satisfied with the Agile approach and speak highly about it.

## **5. CONCLUSION AND FUTURE WORK**

This paper reports our experience in moving from Waterfall process to Agile process in conducting the software engineering capstone project. This move involves restructuring the deliverables and the assessment plans of the two-semester project. To evaluate the impact of the change, a survey was conducted to solicit students' feedback. Twenty-eight students filled the survey. Results show that using Agile helped students in have better design, early training, better risk management and more satisfaction.

In the future offering of the course, we plan to provide more emphases on unit testing and encourage student to use automated unit testing tools such as Junit as currently the focus is more on integration and system testing. Furthermore, we plan to divide the project into three iterations, one in the first semester with 30% of the project features and two in the second semester with 30% and 40% respectively. This decision is based on our observation that more iterations enable the students to have better understanding of the requirements and design and early enhancement of team capabilities by identify training needs. Introducing another iteration will also provide students with opportunity to gain experience in refactoring.

## **ACKNOWLEDGMENTS**

The authors acknowledge the support of King Fahd University of Petroleum and Minerals.

**REFERENCES**

- [1] V. Mahnic, "A Capstone Course on Agile Software Development Using Scrum," *IEEE Transactions on Education*, vol. 55, no. 1, pp. 99-106, 2012.
- [2] M. I. Alfonso and A. Botia, "An Iterative and Agile Process Model for Teaching Software Engineering," in *18th Conference on Software Engineering Education & Training (CSEET'05)*, 2005, pp. 9-16.
- [3] B. Lu and T. DeClue, "Teaching agile methodology in a software engineering capstone course," *J. Comput. Sci. Coll.*, vol. 26, no. 5, pp. 293-299, 2011.
- [4] K. Keefe and M. Dick, "Using Extreme Programming in a capstone project," presented at the *Proceedings of the Sixth Australasian Conference on Computing Education - Volume 30*, Dunedin, New Zealand, 2004.
- [5] T. Smith, K. M. L. Cooper, and C. S. Longstreet, "Software engineering senior design course: experiences with agile game development in a capstone project," presented at the *Proceedings of the 1st International Workshop on Games and Software Engineering*, Waikiki, Honolulu, HI, USA, 2011.
- [6] D. Rover, C. Ullerich, R. Scheel, J. Wegter, and C. Whipple, "Advantages of agile methodologies for software and product development in a capstone design project," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, 2014, pp. 1-9.
- [7] J. G. Kuhl, "Incorporation of Agile Development Methodology into a Capstone Software Engineering Project Course," in *The 2014 ASEE North Midwest Section Conference*, 2014, pp. 1-8.
- [8] D. Ding, M. Yousef, and X. Yue, "A case study for teaching students agile and scrum in Capstone course," *J. Comput. Sci. Coll.*, vol. 32, no. 5, pp. 95-101, 2017.
- [9] C. Coupal and K. Boechler, "Introducing agile into a software development Capstone project," in *Agile Development Conference (ADC'05)*, 2005, pp. 289-297.
- [10] V. Devedzic and S. R. Milenkovic, "Teaching Agile Software Development: A Case Study," *IEEE Trans. on Educ.*, vol. 54, no. 2, pp. 273-278, 2011.
- [11] D. F. Rico and H. H. Sayani, "Use of Agile Methods in Software Engineering Education," in *2009 Agile Conference*, 2009, pp. 174-179.
- [12] D. Knudson and A. Radermacher, "Updating CS capstone projects to incorporate new agile methodologies used in industry," in *2011 24th IEEE-CS Conference on Software Engineering Education and Training (CSEE&T)*, 2011, pp. 444-448.
- [13] R. T. d. Souza, S. D. Zorzo, and D. A. d. Silva, "Evaluating capstone project through flexible and collaborative use of Scrum framework," in *2015 IEEE Frontiers in Education Conference (FIE)*, 2015, pp. 1-7.
- [14] A. El-Abbassy, R. Muawad, and A. Gaber, "Evaluating Agile Principles in CS Education," *International Journal of Computer Science and Network Security*, vol. 10, no. 10, pp. 19-29.

# 4D AUTOMATIC LIP-READING FOR SPEAKER'S FACE IDENTIFICATION

Adil Abdulhur AboShana

Department of Information System,  
University of eötvös loránd, Budapest, Hungary

## ABSTRACT

*A novel based a trajectory-guided, concatenating approach for synthesizing high-quality image real sample renders video is proposed. The lips reading automated is seeking for modeled the closest real image sample sequence preserve in the library under the data video to the HMM predicted trajectory. The object trajectory is modeled obtained by projecting the face patterns into an KDA feature space is estimated. The approach for speaker's face identification by using synthesise the identity surface of a subject face from a small sample of patterns which sparsely each the view sphere. An KDA algorithm use to the Lip-reading image is discrimination, after that work consisted of in the low dimensional for the fundamental lip features vector is reduced by using the 2D-DCT. The mouth of the set area dimensionality is ordered by a normally reduction base on the PCA to obtain the Eigen lips approach, their proposed approach by[33]. The subjective performance results of the cost function under the automatic lips reading modeled, which wasn't illustrate the superior performance of the method.*

## KEYWORDS

*Lip Segmentation, Discrete cosine transform algorithm, kernel Discriminant Analysis, Discrete Hartley Transform, hidden Markov Model.*

## 1. INTRODUCTION

It represented once for any viseme image visible and model 'visual silence'. The hidden Markov model(HMM) is trained from the visual features with three states and a diagonal covariance Gaussian Mixture Model (GMM) associated with each task are sequences of embedded trained and test with view angle dependence. The approach was visual speech synthesis and the visual parameters were generated from HMM by using the dynamic ("delta") constraints of the features. The mouth motion under the video can be rendered from the predicted visual parameter trajectories. The drawback of the HMM-based visual speech synthesis method is generated blurring due to feature dimension reduction in statistical modeling, i.e. PCA and the maximum likelihood(MLL). Proposed by [5][6].

## 2. RELATED WORK

The employed lip-reading systems uses developed features, i.e. csamandhilda, that are consistent across a view, which indicates improved robustness to viewpoint updated for all the types of the primitive feature. The goal of this experiment is to obtain the best viewing angle for computing lip-reading and active appearance model (AAM) features that are extracted from each view, respectively by using their second derivatives. They use a linear predictor based on the tracking and it's has more robust lip-contour than the AAM That was introduced by [9]. The audio-visual speech recognition system, visual features obtained from Discrete Cosine Transform (DCT) and active appearance model. (AAM) were projected onto a 41 dimensional feature space using the LDA, proposed by [34]. The systems reduce the dimensionality for Linear Discriminant Analysis (LDA) or Fisher's Linear Discriminant (FLD) as introduced by [57]. The project pixel and the colour information pixels both are using a lower dimensional space. The threshold operation based on the lower dimensional space is represented for the lip segmentation. [58] located the face region with a skin-colour model. The mouth region was localized by involving in the skin region.

The lips region segmentation was using the G and B components of the RGB colour based contented for Fisher transform vectors. Adaptive thresholds representation as an operation of the grey scale histogram of the image were then employed to segment the lip pixels. [59] used RGB values from training images to learn the Fisher discriminant axis. The mouth region colour onto their axes used to enhance the lip-skin boundary then a threshold was applied to segment the lip pixels. [60] used an identical approach to lip segmentation. [41] used FLD by visible for an edge detection, dynamic based on the split ability of multi-dimensional distributions. The edge of the lip contour extraction was described as the point at which there is a maximized distributions of lip and non-lip pixels. Principal Components Analysis (PCA), as suggested by [57] has been reduced to dimensionality automatic for the identification of lip pixels.

### 2.1 Segmentation

Pixel-based of any lip segmentation have been using a specific colour space appearance to compare involve for pixel colour combined with a set of thresholds. They changed image from the binary separation into a part of lip pixels and non-lip pixels. They used colour spaces by choice and the operation threshold base by histogram selection. [50] worked a normalized RGB colour which is created based on a process to satisfy the maximum intensity normalization. They had represented colour components based on thresholding was beneficial to classify the image into lip and non-lip pixels. [51] represented a mouth region by using three different colour appearance in YCbCr, RGB and HSV. Lip pixels segmentation obtained by thresholding the YCbCr, Green and Hue components of the colour appearance and associated the results by using a logical and operator.

Region based lip Segmentation systems use developed cost-functions to constrain the subset of pixels being chosen. In region based lip segmentation these cost functions usually include the shape constraints or the locality of lip pixels. The aim of this section is to use some criterion functions to choose chromatically and homogeneous pixels in an image to be the lip region. The approach to region-based lip segmentation is the Markov Random Fields technique (MRF), proposed by [51]. Each lip pixel is processed as a stochastic variable sensitive by a result of exactly the neighborhood is connected. The identification of an objects in an image formula

where each object referred to as a set is the like to the a single MRF. In[53],use a spatiotemporal neighborhood compute to form a region lip segmentation using MRF. The temporal input is the difference between the binary labels in two consecutive images, respectively [54] use primitive pixel dependent on thresholding operation under the HSV mouth region image with edge information to create a label the MRF set as part of the lip or non-lip regions. They weren't using spatial homogeneity, pixel-based algorithms attempted to be faster than region-based approaches from time to time generating coarse segmentation results and no re- correctly classify for pixels is noisy. The pixel appearance is relative to the colour as define and the local neighborhoods of pixels was included. Lip pixels segmentation were practiced by a Bayesian classifier ,this introduced by [56] which uses Gaussian Mixture Models (GMM) that are estimated by using the Expectation Maximization algorithm (EM) , refer to [55] and their normalized the R and G components of the RGB space by using the intensity of the pixel. The normalization was an illumination invariant to RGB. Lip pixels segmentation were used by a thresholdin colourspace.

## 2.2 Data Acquisition

The dataset was used to record and study by [7].The datasets contain discrete English phonemes correspond to the visemes visible in The face . The face model of MPEG-4 standards is used two Facial Animation Parameters and Facial Definition Parameters. The visemes connected to form words and sentences is due the specification of used visemes as the recognition unit. The calculated number of visemes is less than phonemes , due speech is partially visible, refer to [8].Video data was recorded by a movie camera in a typical mobile environment. The camera view was on the speaker's lips-reading and it was also kept fixed throughout the recordings. Factors, examples window size (240x320 pixels), view angle of the camera, background and illumination were kept constant for each speaker. To validate the proposed method, 12 subjects (6males 6 females) were used. Each speaker, recorded 6 phonemes at a sampling rate of 30 frames/sec and every phoneme was recorded five more times to give a sufficient variability.

## 3. PROPOSED METHOD

They used automated lip-reading, consist of 2D DCT and Eigen-lips. The lip shape of an AAM algorithm defined by,  $s$ , is associated with coordinates  $(x,y)$  of the set  $N$  vertices that determined the features on an object:  $s = (x_1, y_1, \dots, x_n, y_n)^T$ . A model that appeared a linear variation in the shape is explain in below equation,

$$S = s_0 + \sum_{i=1}^m p_i s_i \quad (1)$$

Where  $s_0$  is a term called mean shape and  $s_i$  are defined by the eigenvectors corresponding to the largest number ( $m$ ) of the covariance matrix consist of the eigenvectors. The coefficients ( $p_i$ ) are defined for the shape parameters that associated with each eigenvector in the shape ( $s$ ) is appeared. The model is always calculated by using Principal Component Analysis (PCA) to a set of shapes handle in a corresponding for each image. To get the shape vertices  $s$ , required in Equation (1) and to compute the shape parameters, it was proposed by [12]. The mouth of the set area dimensionality is ordered by a normally reduction base on the PCA to obtain the Eigenlips approach, their proposed approach by[33].

The area (A), a term of an AAM is represented by the pixels that stretch inside the mesh  $s_0$ . AAMs represented as linear a variation visible, so is appeared as term base on  $A_0$  plus a linear associated with to display the images  $A_i$ .

$$A = A_0 + \sum_{i=1}^l \lambda_i A_i \quad (2)$$

where  $\lambda_i$  are represented as parameters. As well, shapes  $s$ , represented the base on  $A_0$  is the mean shape normalized image and vectors  $A_i$  are the reconstructed the shaped eigenvectors corresponding to the largest eigen values and both are always calculated by using PCA to the shape training images, it is normalized by [11]. The scenario, established a lipin contained on the speaker's face identification system based on optimal performance of the control. A lip is necessary for discrimination, after that the their work consisted of in the low dimensional for the fundamental lip feature vector and reduced by using the 2D-DCT. They found that subjects have a two-stage discrimination analysis for speaker identify, such as to exploit two pair correlations temporal and spatial correlations by [16]. The eigen face approaches [18][17] are using the principal component analysis (PCA), or Karhunen-Loeve transforms (KLT). It obtained to account the statistical base on the measure between the pixel values of images to be visual in a training update to create an orthogonal for representing images. The eigenvectors of the covariance matrix of the training update of face images are calculated and they are retrained for describing the images are called eigen faces. The ones corresponding to the largest eigen values of the covariance matrix for each training and test face is characterized by its projection on the eigen faces, and the comparison of two faces is obtained by comparing two sets of projections.

### 3.1 Lips Transformation

The maximum flexibility in deforming to movie shapes of body likely lips shapes .so that, then applying for complex shapes must have control points to describe them, this is usually implied, to instability in tracking [24][23]. A movie shapes, provided the lips are not flexing, is an approximately rigid, planar shape under orthographic projection [25]. They show that form a vector  $Q$  for the lips a represented body with the affine transformation that is applies to the template to get the shape. No-rigid motion can be handled in the same way, by providing that represents both the rigid and non-rigid degrees of freedom of the shape. This is used, for instance, to figure out the movements in hand tracking, or to automatic lip movements. It is crucial point therefore to allow the right degrees of freedom for non-rigidity, neither too few resulting in excessive stiffness, nor leading to instability and then the snake-based methods for handling non-rigid motion by [22] allowed for the high degrees of freedom which leads to instability, well unusable for real-time. Their scenario, the framework of localized color active contour model (LCACM) expand from scheme, introduced by [19] given that the foreground and background regions with variation in color space. They utilize a 16-point deformable model by [20] with geometric constraints to achieve lip contour extraction. They used deform modelled location region base to approach lip tracking combined with the extraction of lips contour in the color image [21].

The dimensionality is reduced for colour-space transforms, generally hue-based transforms are used in pixel based lip segmentation systems. [54] were changing the red hue, value base on a threshold to identify the lip pixels. The hue transforms originally, proposed by [61]. Their purpose is to transform the colour space to maximize the chromatic difference between skin and lips that led to use based on the connected between the R and G components of the RGB system.



The transform is combined with pixel-based classification by [62]. In [63], used to transform the information on the RGB colour under the CIELUV space. The adaptive used histogram-based operation thresholds later resulted in a binary classifier for lip segmentation.

Discrete Hartley Transform( DHT) is subjected to wavelet multi-scale edge detection to obtain the lip contour which is smooth by using morphological operations, referred [42] obtain in the mouth region by using generally directed camera and subject of the region hue colour transform. hybrid of representing edge which are used edge enhancement exactly a polynomial curve, suggested by [49] for the mouth region to the YCbCr colour space transformation. A parametric model using cubic curves is used to detect the lip contour. [48] approach with normalised RGB values as the colour features. A geometric model is used to detect the lip contour. In [47], use a cost function that the boundary of the mouth region is closed or open combined with a parabolic curve to extract the lip contour. [46] employed lip contour extraction using cubic B-Splines to boundary the lip-pixels extracted, however, the classification is a binary processed. Snakes for segmentation were proposed by [45]. Active contours give a deformable model of a contour and it is an object under image by using internal and external cost functions or energy functions to lead the model to satisfy the object boundary. The idea of using “hybrid edges” was extended with a snakes formulation is “jumping” defined to extract the lip contour by [43]. [44] use an active contour style energy design to detect the inner lip contour. In [42], use active contours to build a geometric template with a mouth region image, proceeded that the keypoints of the lip contour are using pixel extracted. B-Spline based active contour is proposed by [41]. [40] use an active contour based codebook to synthesize similarly lip contours under any image. The convergence of active contour created is defined in the Gradient Vector Flow (GVF), introduced by Xu and Prince. (1997) uses edge represented and an edge-based vector-field to formulate that can be external energy for snakes. GVF snakes are used for lip contour detection are proposed by [64]. Active Shape Models (ASM) [38] and [28] used Active Appearance Models (AAM) and they have been used for contour-based lip segmentation. The active contours mod-based approach to lip segmentation uses level sets. It provides an energy minimization by curve created called B-Spline technique that has been created based on Widely model shapes and object boundaries in the computer vision presented by [37].

### 3.2 Synthesizing Identity Surfaces

In [27][28][29], use analysis base on synthesis. They approach by using the synthesise identity surface of a subject face from a small sample of patterns which sparsely fill the view sphere. The base, approximate of the identity surface using a set of  $N_p$  planes separated by  $N_p$  multiple views. They used PQI tilt and yaw are the  $z$  discriminating feature vector of a face pattern.

KDA vector.  $(x_{01}, y_{01}), (x_{02}, y_{02}), \dots, (x_{0N}, y_{0N})$  are define views which separate the view plane into  $N_p$  spieces. On each of these  $N_p$  spieces, the identity surface is approximated by a plane suppose the  $M_i$  sample patterns filled by the plane are  $(x_{01}, y_{01}, z_{01}), (x_{02}, y_{02}, z_{02}), \dots, (x_{0M}, y_{0M}, z_{0M})$ . This is a quadratic big problem which can be solved using the interior point method by [26]. They can classify the pattern into one of the face classes by computing the distance to all of the identity surfaces as the Euclidean distance between  $z_0$  and the corresponding point on the identity surface called  $(z)$ .

$$d = \|z_0 - z\| \quad (3)$$

An object trajectory is obtained by projecting the face patterns into the KDA feature space. In same time , according to the pose information about the face patterns. They can build the model trajectory on the identity surface of each subject using the same pose information and temporal order of the object trajectory. Those two kinds of trajectories, i.e. object and model trajectories, encode the spatio-temporal information on the tracked face. Hence , the recognition problem can be solved by matching the object trajectory to register for the set of model trajectories. The primitive achievement of trajectory matching face is applying by computing the trajectory distances, it reaches to the time of the frame called (t).

$$d_m = \sum_{i=1}^t w_i d_{mi} \quad (4)$$

where d, the pattern distance between the face pattern catches in the frame and the identity surface of the subject, is computed from (3), and  $(w_i)$  term is the weight on this distance.

### 3.3 Trajectory for sequence position lips

The novel trajectory Lead to the lips sample selection approach is proposed. In training, the image samples are sequences(S) encoded in low-dimensional visual feature vector. The feature vector is used to train HMM trajectory  $\lambda$  model that is a statistical model. The trained model gives the best feature trajectory by using a maximum likelihood (MLL) that is sensitive. The last status is to reconstruct the optimal feature trajectory drawback by  $\bar{S}$  term in the original high-dimensional sample space. The low-dimensional visual parameter trajectory to samples in the sample space. In implementing used the HMM lead to predicted trajectory  $\bar{V}$ , a smooth image sample sequence  $\bar{S}$  is sought more best from the sample library and the mouth sequence is then returned back to a background primitive recorder for video. The lips images, has a large number of the Eigen lips contained of the accumulated variance. The visual feature of each lips image is formed by its PCA vector,  $V^T = s^T w$  where  $w$  is the projection matrix made by number Eigen lips. We use the specially algorithm to specify to the best visual parameter vector sequence.

$$V = [V_1^T, V_2^T, \dots, V_T^T]^T \quad (5)$$

By giving maximization for the maximum likelihood (LM) algorithm . The HMM predicted visual parameter trajectory had detailed to move a compact description , in the lower level eigen-lips space. However, the lips image sequence shown at the top of is blurred due to dimensionality reduction in PCA and MLL-based model parameter estimation and trajectory is obtained . To solve this blurring, suggest the trajectory is leading to real sample sequence approach to constructing from. Hence , the detailed movement in the visual trajectory is reconstructed and image real sample rendering is truth. This was propose [30].The unit obtained in concatenative speech synthesis , the cost count for a sequence of trajectory called  $T$  choice samples are the weighted sum of the target and concatenation costs:

$$C(\hat{v}_1^T, \hat{s}_1^T) = \sum_{i=1}^T \omega^t c^t(\hat{v}_i, \hat{s}_i) + \sum_{i=2}^T \omega^c c^c(\hat{v}_{i-1}, \hat{s}_i) \quad (6)$$

The target cost of an image sample (s) is dependent over the measured of the Euclidean distance between their PCA vectors.

The concatenation cost is measured based on the normalized 2-D cross correlation (NCC) between two image samples  $\hat{s}_i$  and  $\hat{s}_j$ . Since the correlation coefficient ranges in value from -1.0 to 1.0, NCC is in nature a normalized similarity score, proposed by [1].

$$c^t(\hat{v}_i, \hat{s}_i) = \|\hat{v}_i - \hat{s}_i^T\| \quad (7)$$

## 4. EXPERIMENT AND RESULTS

### 4.1 Expire Vector Feature

They created separate shape and appearance model to encode any view independently, hence the shape feature (parameter,  $p$ ), and app feature (parameter  $\lambda$ ), this introduced by [10]. They need for two phased for associated the shape and appearance parameters. First, they used primitive way by combining the feature vectors (CFV), which called a cat and the second is concatenating the features and reduce the dimensionality using PCA, proposed by [14] which used (csam)/feature. CFV improved by using an LDA over window for the set of frames. It is proposed by [15] which went to represented (hilda) features in the frontal lip-reading. It has applied for two features are the discriminating, it is introduced by [9] For all features, a z-score normalization is used, which has been visible to develop the separability between the features of the classes by [31]. The best viewing angle for the primitive features, i.e., those that aren't relative to a third PCA or an LDA i.e (shape, app and cat) seems to be more one angle of a view.

### 4.2. Expire for across multiple camera

They used audio-visual speech dataset base on called LiLIR. The dataset contains multi-camera, multi-angle recordings of a speaker recite a lesson 200 sentences from the supplier arrangement Corpus. The structure and size of the LiLIR dataset has enable to train the hidden Markov models (HMMs) onset of the word for visual speech units, such visemes, hence the number words as representative for the vocabulary of the database and it is approximately 1000 words, datasets used to satisfy automatic lip-reading for the each view camera. The dataset contains multiple type of camera such as two HD cameras recording. As well, there are three SD camera sand 60 viewpoints. All cameras were synchronization locked during recording, proposed by [13][32].

### 4.3. Expire Result

In [36], used the mouth cavity region implemented in a banalization process, As well, a time change of the two features is expressed as a two-dimensional a trajectory of the lip motion of the target word. Observing these lip images, in the table .1, it can show us the visual speechless person's male concept that 30th frame lip image expressed the shape of "Stick", 40st, 30th, and 50th frames expressed "cake", "torsion", "two", "with", and "under", respectively. in the table .2, it can show us the visual speechless person's female concept, that 30th frame lip image expressed the shape of "Stick", 40st, and 50th frames expressed "cake", "Torsion", "two", "with", and "Under", respectively.

Proposed by [35], recorded 50 times for each word, and recorded 500 image sequence one subject. The image size is 320×240 pixels and the frame rate is 30 frame/sec. The lip detection

was applying for 500 image sequences. Time passes along the direction of the arrow base on the automatic of the lip in an utterance. It can show us that trajectory in all words is being drawn by the visual observation. The recognition process was applying with feature sets. Figure.1 shows a trajectory which the word is "a stick", "torsion", "cake", "under", "with" and "two", respectively. The horizontal axis is area  $s$ , and the vertical axis is an aspect ratio area ( $A$ ). Plotted circle marks in this figure are the position of vector features of all the frame, and these marks are connected with a line.

Table 1. Illustrating Visual Speech Person's Female Concept


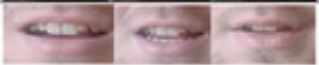
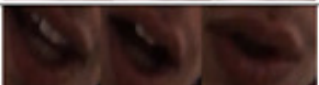
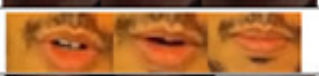




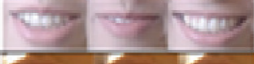

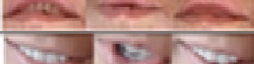

Month	Words	Production	visemes	Gander
January	A Stick	/a/ /s/ /t/ /ɪk/		Female
March	Two	/t/ /u/		Female
May	Cake	/c/ /a/ /ke/		Female
September	under	/u/ /n/ /d/		Female
October	Torsion	/t/ /oʊ/ /sion/		
December	with	/w/ /ɪ/ /th/		Female

Table .2.Illustrating Visual Speech Person's Male Concept

days	words	Phonemes	Vismenes	Gander
1	A stick	/a/ /s/ /t/ /ɪk/		Male
2	cake	/c/ /a/ /k/		Male
3	Under	/u/ /n/ /d/		Male
4	Two	/t/ /u/		Male
5	Torsion	/t/ /oʊ/ /sion/		Male
	with	/w/ /ɪ/ /th/		Male

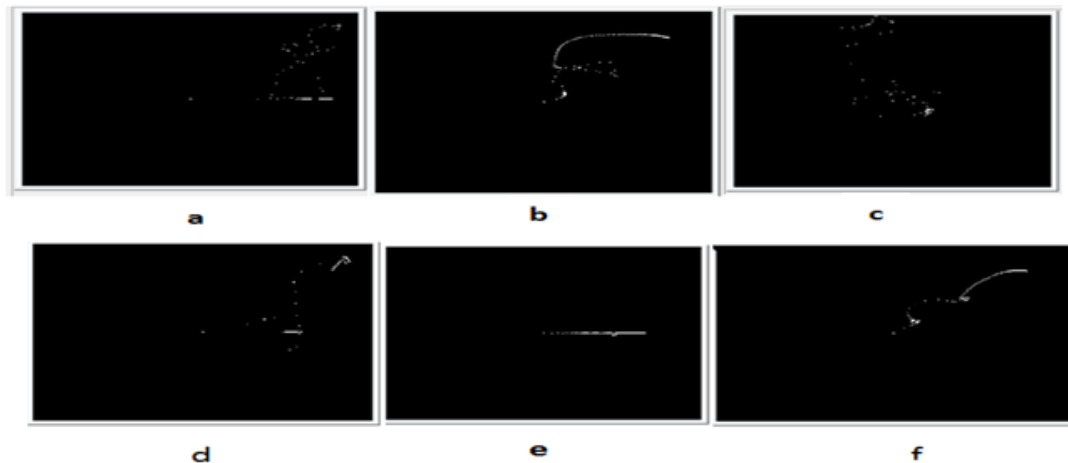


Figure.1. Image of computed trajectory and correspond Lip

## 5. CONCLUSION

We propose a trajectory-guided, real sample concatenating approach for synthesizing high-quality automatic image-real articulator. Objectively, we evaluated the performance of our system in terms of speaker's face identification by using automatic lip reading represented in the visual domain. The system framework using the signature of the visemes approaches by track trajectory for lip contour extraction as represented the whole word. The target word is recognized based on the word's English included two types of gender female and males were shown that recognition using the trajectory vector feature is obtained the vocabulary of the database is approximately more than 100 words, data sets used to satisfy the automatic lip-reading across multi-view camera.

## REFERENCES

- [1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a largespeech database," Proc. ICASSP 1996, pp. 373-376.
- [2] D. Sweet's, J. Weng, Using discriminant eigen features for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (1996) 831-836.
- [3] B. Scholkopf, A. Smola, K.-R. Muller, Kernel principal component analysis, in: W. Gerstner, A. Germond, M. Hasler, J.-D. Nicoud (Eds.), Artificial Neural Networks—ICANN'97, Lecture Notes in Computer Science, Springer, Berlin, 1997, pp. 583-588.
- [4] V. Roth, V. S. Solla, T. Leen, K.-R. Mu'ller, "Steinhage, Nonlinear discriminant analysis using kernel functions", in: (Eds.), Advances in Neural Information Processing Systems 12, MIT Press, Cambridge, MA, 1999, pp. 568-574.
- [5] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based Text-To-Audio-Visual Speech Synthesis," ICSLP 2000.
- [6] L. Xie, Z.Q. Liu, "Speech Animation Using Coupled Hidden Markov Models," Pro. ICPR'06, August 2006, pp. 1128-1131
- [7] W. C. Yau, D. K. Kumar, and S. P. Arjunan. "Visual speechrecognition using dynamic features and support vectormachines, International Journal of Image and Graphics, vol.8, pp. 419-437, 2008.
- [8] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," IEEE Transactions on Speech and Audio Processing, 14(3), (2006), 1082-1089.
- [9] E. Ong, Y. Lan, B. Theobald, H. R., and R. Bowden, "Robust facial feature tracking using selected multi-resolution linear predictors," in Proc. of ICCV, 2009.

- [10] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, pp. 796–804, 1968.
- [11] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.
- [12] Y. Lan, R. Harvey, B. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lip-reading," in *Proceedings of Proceedings of International Conference on Auditory-Visual Speech Processing*, 2009, pp. 102–106.
- [13] Y. Lan, B. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, "Improving visual features for lip-reading," in *Proceedings of Proceedings of International Conference on Auditory-Visual Speech Processing*, 2010.
- [14] T. Cootes and C. Taylor, "Statistical models of appearance for computer vision," *Imaging Science and Biomedical Engineering*, University of Manchester, Tech. Rep., 2004.
- [15] G. Potamianos, C. Neti, J. Luetttin, and I. Matthews, "Audiovisual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. MIT Press, 2004.
- [16] Cetin Gul, H.E. Yemez, Y. Erzin, E. Tekalp, A.M., "Discriminative lip-motion features for biometric speaker identification," in *IEEE ICIP*, 2004, vol.3, pp.2023–2026.
- [17] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Amer., A*, vol. 4, pp. 519–524, 1987.
- [18] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, pp.71–86, 1991.
- [19] S. Lankton, A. Tannenbaum, Localizing region-based active contours, *IEEE Transactions on Image Processing* 17 (11) (2008) 2029–2039.
- [20] S. Wang, W. Lau, S. Leung, Automatic lip contour extraction from color images, *Pattern Recognition* 37 7 (12) (2004) 2375–2387.
- [21] Y.M.Cheung , X.Liu , X. You .A local region based approach to lip tracking, *Int. Journal Pattern Recognition*, Vol. 45 Iss. 9 pages 3336–3347, Sep.2012 .
- [22] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. In *Proc. 1st Int. Conf. on Computer Vision*, pages 259–268, 1987.
- [23] J.J. Koenderink and A.J. Van Doorn. Affine structure from motion. *J. Optical Soc. of America A.*, 8(2):337–385, 1991.
- [24] A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *Int. Journal of Computer Vision*, 11(2):127–145, 1993.
- [25] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.
- [26] R. Vanderbei. Loqo: An interior point code for quadratic programming. Technical report ,Princeton University, 1994. Technical Report SOR 94-15.
- [27] Ezzat, T. and Poggio, T. 1996. Facial analysis and synthesis using image-based methods. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Vermont,US, pp. 116–121.
- [28] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *European Conference on Computer Vision*, volume 2, pages 484–498, Freiburg, Germany, 1998.
- [29] T.Vetter,. and V.Blanz, . 1998. Generalization to novel views from a single face image. In *Face Recognition: From Theory to Applications*, (Eds.), Springer-Verlag, pp. 310–326.
- [30] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE. Trans. Computers*, C-22(1), 1973.
- [31] Y. Lan, B. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, "Improving visual features for Lip-reading," in *Proceedings of Proceedings of International Conference on Auditory-Visual Speech Processing*, 2010
- [32] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "Resource management RM2 2.0," *Linguistic Data Consortium*, Philadelphia, 1993.
- [33] C.Bregler., Y.Konig., (1994) "Eigenlips For Robust Speech Recognition", *Proc.of ICASSP'94*, Vol. II, Adelaide, Australia, p669-672.
- [34] C.Neti ., G. Potamianos., J.Luetttin., (2000) "Audio-visual speech recognition", *Final Workshop 2000 Report*, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD.
- [35] T. Saitoh and R. Konishi, "Lip reading based on sampled active contour model," *LNCS3656*, pp. 507–515, September 2005.

- [36] M. J. Lyons, C.-H. Chan, and N. Tetsutani, "Mouth Type: text entry by hand and mouth," *Proc. of Conference on Human Factors in Computing Systems*, pp. 1383-1386, 2004.
- [37] A. Khan, W. Christmas, and J. Kittler. Lip contour segmentation using kernel methods and level sets. In *ICVS*, volume 4842:II, pages 86–95, 2007.
- [38] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models -their training and application. *CVIU*, 61(8):36–59, 1995.
- [39] C. Xu and J.L. Prince. Gradient vector flow: A new external force for snakes. In *CVPR*, pages 66 – 71, 1997.
- [40] K.F. Lai, C.M. Ngo, and S. Chan. Tracking of deformable contours by synthesis and match. In *ICIP*, volume 1, pages 657 – 661, 1996.
- [41] T. Wakasugi, M. Nishiura, and K. Fukui. Robust lip contour extraction using separability of multi-dimensional distributions. In *FGR*, pages 415 – 420, 2004.
- [42] P. Delmas, N. Eveno, and M. Li'evin. Towards robust lip tracking. In *ICPR*, 2002.
- [43] N. Eveno, A. Caplier, and P.Y. Coulon. Jumping snakes and parametric model for lip segmentation. In *ICIP*, volume 2, pages 867 – 870, 2003.
- [44] S. Stillitano and A. Caplier. Inner lip segmentation by combining active contours and parametric models. In *VISAPP*, pages 297 – 304, 2008.
- [45] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321 – 331, 1988.
- [46] M. S'anchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking with b-splines. In *AVBPA*, pages 69–76, 1997.
- [47] L. Zhang. Estimation of the mouth features using deformable templates. In *ICIP*, volume 3, pages 328 – 331, 1997.
- [48] S. Werda, W. Mahdi, and A. BenHamadou. Colour and geometric based model for lip segmentation. In *ICIP*, pages 9 – 14, 2007.
- [49] A.E. Salazar, J.E. Hernandez, and F. Prieto. Automatic quantitative mouth shape analysis. *Lecture Notes in Computer Science*, 4673:416–423, 2007.
- [50] J.A. Dargham and A. Chekima. Lips detection in the normalised rgb colour scheme. In *ICTTA*, volume 1, pages 1546 – 1551, 2006.
- [51] E. Gomez, C. M. Travieso, J. C. Briceno, and M. A. Ferrer. Biometric identification system by lip shape. In *ICCS*, pages 39 – 42, 2002.
- [52] H. Bunke and T. Caelli. *Hidden Markov Models: Applications in Computer Vision*. World Scientific Publishing Co., 2001.
- [53] F. Luthon, A. Caplier, and M. Li'evin. Spatiotemporal mrf approach to video segmentation Application to motion detection and lip segmentation. *Signal Processing*, 76(1):61 – 80, 1999.
- [54] X. Zhang and R.M. Mersereau. Lip feature extraction towards an automatic speechreading system. In *ICIP*, volume 3, pages 226 – 229, 2000.
- [55] Y. Nakata and M. Ando. Lipreading methods using color extraction method and eigenspace technique. *Systems and Computers in Japan*, 35(3):1813 – 1822, 2004.
- [56] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, 1990.
- [57] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001.
- [58] J.M. Zhang, D.J. Wang, L.M. Niu, and Y.Z. Zhan. Research and implementation of real time approach to lip detection in video sequences. In *ICMLC*, pages 2795 – 2799, 2003.
- [59] R. Kaucic and A. Blake. Accurate, real-time, unadorned lip tracking. In *ICCV*, pages 370–375, 1998.
- [60] W. Rongben, G. Lie, T. Bingliang, and J. Linsheng. Monitoring mouth movement For driver fatigue or distraction with one camera. In *ITSS*, pages 314–319, 2004.
- [61] A.C. Hulbert and T.A. Poggio. Synthesizing a color algorithm from examples. *Science*, 239(4839):482 – 485, 1988.
- [62] N. Eveno, A. Caplier, and P.Y. Coulon. New color transformation for lips segmentation. *IEEE Fourth Workshop on Multimedia Signal Processing*, pages 3 – 8, 2001.
- [63] Y. Wu, R. Ma, W. Hu, T. Wang, Y. Zhang, J. Cheng, and H. Lu. Robust lip localization using multi-view faces in video. In *ICIP*, pages IV 481 – 484, 2007.
- [64] L.E. Mor'an and R. Pinto. Automatic extraction of the lips shape via statistical lips modelling and chromatic feature. In *CERMA*, pages 241 – 246, 2007.

**AUTHOR**

**Adil A. Aboshana** received the B.S. mathematical from university salah-adeen , Science Iraq in 1985 and 1989 respectively ,and he received M.S. degrees from department of the information Technology, Science, University of Utara –Malaysia ,in 2007 and 2009, respectively. He is currently studying PhD in the department of information system. His research interests include, pattern recognition and image computer vision, processing with applications to biometrics. He is one of the participants who received paper award at the conference IEE in 2015





# ANALYSIS OF LAND SURFACE DEFORMATION GRADIENT BY DINSAR

Karima Hadj-Rabah<sup>1</sup>, Faiza Hocine<sup>2</sup>, Sawsen Belhadj-Aissa<sup>3</sup> and  
Aichouche Belhadj-Aissa<sup>4</sup>

<sup>1,2,3,4</sup>Department of Telecommunications, University of Sciences and  
Technology Houari Boumediene (USTHB), Algiers, Algeria

## ABSTRACT

*The progressive development of Synthetic Aperture Radar (SAR) systems diversify the exploitation of the generated images by these systems in different applications of geoscience. Detection and monitoring surface deformations, procreated by various phenomena had benefited from this evolution and had been realized by interferometry (InSAR) and differential interferometry (DInSAR) techniques. Nevertheless, spatial and temporal decorrelations of the interferometric couples used, limit strongly the precision of analysis results by these techniques. In this context, we propose, in this work, a methodological approach of surface deformation detection and analysis by differential interferograms to show the limits of this technique according to noise quality and level. The detectability model is generated from the deformation signatures, by simulating a linear fault merged to the images couples of ERS1 / ERS2 sensors acquired in a region of the Algerian south.*

## KEYWORDS

*Radar Interferometry (InSAR & DInSAR), Surface Deformation Gradient (SDG), Coherence, Resolution, Filtering*

## 1. INTRODUCTION

The interferometry (InSAR) is a technique allowing the generation of altimetric information and its variations from couples of SAR radar images. Although, interferometric techniques had known an important development in terms of correction treatments and telemetric phase analysis approaches and methods, their evaluation in comparison with the required precision by geodesy, showed limitations as for their use in cartography deformation and sources characterization causing these last. The quality of interferometric products depends, from one side, on the acquisition geometry and periodicity of radar systems, and from the other side on, the atmospheric conditions, the scene degradations and the surface state observed at a moment 't'. As a result, differential interferometry (DInSAR) processes do not allow in all cases the deformations detection. However, it is essential to know the deformations that can be detected by the differential interferometry, thus making it possible to decide on the best solution to opt for measuring and monitoring a surface displacement phenomenon [1][2].

Massonnet and Feigl [3] had shown that the necessary condition for the detection of a deformation is that the maximum surface deformation gradient (SDG) equals a fringe by the pixel resolution. Indeed, a fringe of deformation equals half a wavelength. This definition depends on two parameters: the wavelength and the resolution, these last are fixed and specified for each

sensor. However, in reference to this relation, SAR radars can detect several types of variable intensity deformations. Unfortunately, in practice, the interferometric measurements are too noisy. These phenomena of noise and phase discontinuity are essentially due to: spatial and temporal decorrelation, atmospheric heterogeneity and others [4][5]. Therefore, the small and the large deformations become undetectable when the level of noise is high.

Otherwise, the estimation of the maximum surface deformation gradient (SDG) does not take into account the noise factor [6] where the first indicator is the interferometric coherence. In practice, the lower the coherence is, the more the maximum SDG decreases [7]. In this context, the work that we are about to present is a contribution to the modelling of the surface deformation gradient from the differentials interferograms. To show the effectiveness of this modelling, we proceeded at first by simulating a brittle deformation (linear fault) which we had integrated in pairs of the ERS1/ERS2 sensor images. The remainder of the paper is organized as follows: a synthesis on the deformation detectability models by DInSAR will be presented in section 2, section 3 will include a description of the methodological approach that we propose. Section 4 will be devoted to the deformation gradient modeling as well as the obtained results. To finish, we generalize in section 5 our modelling for a case of a real deformation. The paper will end with a general conclusion and perspectives in order to enrich this work.

## 2. RELATED WORK

The first formula based on the consideration of Massonet and Feigl [3] defines the maximum deformation gradient as:

$$D_m = \varphi_{2\pi}^{dif} * \frac{1}{r_g} \quad (1)$$

Where:  $\varphi_{2\pi}^{dif}$  is the differential phase of one fringe (a turn of  $2\pi$ ) which corresponds to half the wavelength  $\lambda$  and the pixel resolution  $r_g$ .

In order to introduce the coherence which best describes the noise effect, Baran et al. [6] have proposed a functional model to determine the minimum and the maximum surface deformation gradient for different values of the coherence, by using real images with a simulated deformation (surface subsidence) for a number of looks  $L$  equals 5. Subsequently, Jiang and Li [8] have extended this model in order to adapt it for different number of looks ( $L= 1, 5$ , and  $20$ ). In the same context, Wang and Li [9] had resumed the same work for the acquired data by the PALSAR sensor. Recently, Hadj-rabah and Hocine [10] had proposed an automatic approach for surface deformation detection based on a multi-scales analysis that can be adapted and implemented to determine the minimum and the maximum surface deformation gradient instead of the subjective approach implemented in the works mentioned above. These works are focused on a single type of deformation (mining subsidence) and do not include a spatial filtering step in advance. For our part, we proposed a methodological approach (see Fig. 1) for surface deformation gradient detection and analysis for different resolutions and noise levels. In addition, we have analyzed the deformation detectability for a linear fault.

## 3. METHODOLOGICAL APPROACH

In order to achieve our goal, A linear fault of the surface is modeled by simulating deformation images with different parameters representing its temporal evolution (see Fig. 2). These deformations are then wrapped (converted) into phase images. They have been merged with Single Look Complex (SLC) images considered as the 'master' images giving birth to the modified SLC images containing the deformation signature. The generating process of differential interferograms is, then applied. In the course of this procedure, a step of spatial

filtering is performed and the resolution is gradually changed in order to observe the effect of noise as well as the level of detail on the deformations detectability. the coherence  $\gamma$  and the deformation gradient are calculated for each generated differential interferogram.

Finally, the modeling is carried out basically, on a step of decision. This last consists in deciding if the deformation is detectable or not, interactively.

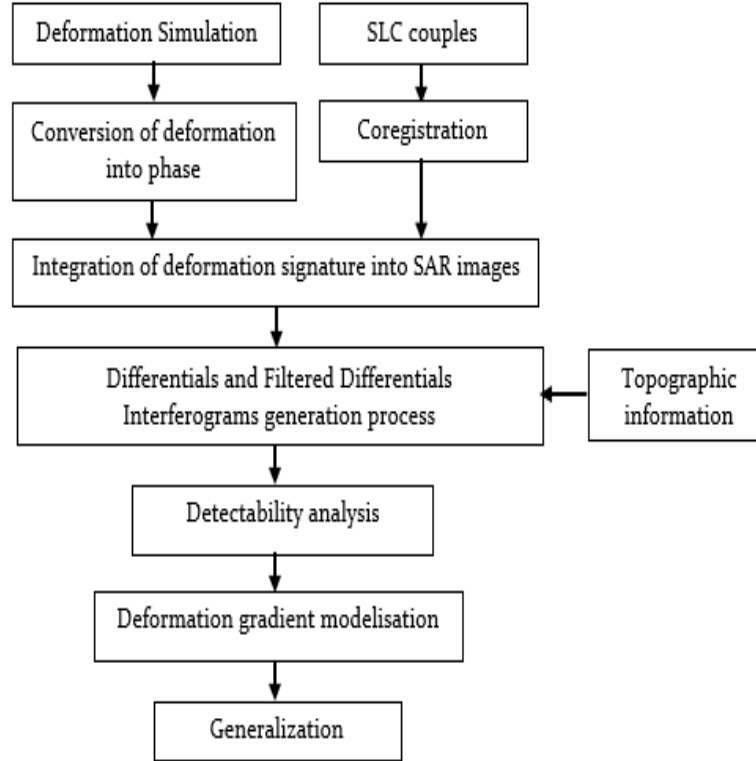


Figure 1. Proposed approach

### 3.1. Deformation Simulation

The simulated deformation according to a direction, supposed to be the acquisition one, is generated by a linear fault. The spatial two-dimensional function  $f$  (following azimuth and distance directions) has been adopted to model such a deformation with a scaling obtained by varying the parameter ' $h$ ', also called the deformation temporal amplitude. Its expression is as follows:

$$f(R, A, h) = h * [A * \cos \theta + R * \sin \theta] \quad (2)$$

With:  $R$  and  $A$ , the coordinates of the deformation image following the range and the azimuthal directions respectively. The value of the angle  $\theta$  representing the fracture line orientation is considered constant, assuming that the deformation coincides with the radar sight angle, thus we obtain a series of simulated images, which their 3D representation shows the deformation evolution as a function of time (see Fig. 2).

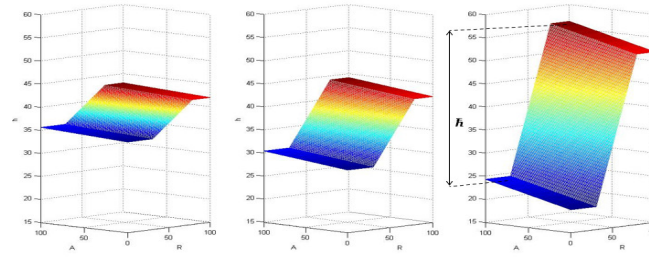


Figure 2. Simulated deformation in 3D

The deformation parameters in the simulated images are represented in the following table:

Table 1. Parameters of simulated deformation.

Name	Simulated Surface Deformation		
	SD1	SD2	SD3
$\theta$	35°		
$h(m)$	0.014	0.084	0.14

### 3.2. Deformation Signature Integration

The variation of the parameter 'h: fault depth' makes it possible to generate three deformation forms at surface, noted SD1, SD2 and SD3. Surface fault signatures are then generated by merging the deformation to one of the two SLC images of the interferometric pairs. The interferograms are generated from the modified images. The selected SLC images have different coherences and resolutions values (8 m, 20 m and 40 m respectively). A filtering step is applied on the complex interferograms consisting to attenuate the noise effect mainly due to spatial decorrelation, geometry effects of acquisition and observed field. Interferometric pretreatments are applied, namely: orbital fringes removal, topographic pair generation from a digital elevation model (DEM) of the region by bringing it back to interferometric pairs geometry.

### 3.3. Detectability Analysis

The deformation fringes in the differential interferograms are analyzed by an interactive method based on rows and columns profile plots. Three groups of interferograms are represented in figures 3 to 6, in order to illustrate the necessary fringes analysis for detectability decision, for different coherences  $\gamma$ , and resolutions values as well as filtering level.

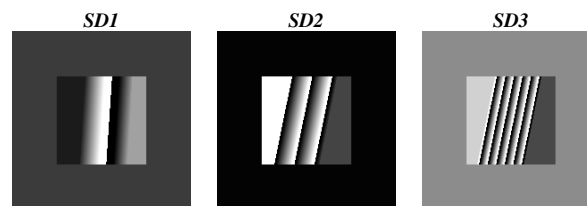


Figure 3. Interferometric phase of the simulated deformations

Fig. 3 and Fig. 4 show the interferometric phase of the simulated deformations (SD1 to SD3) and its corresponding differentials interferograms, with different coherence values for a 20 m resolution.

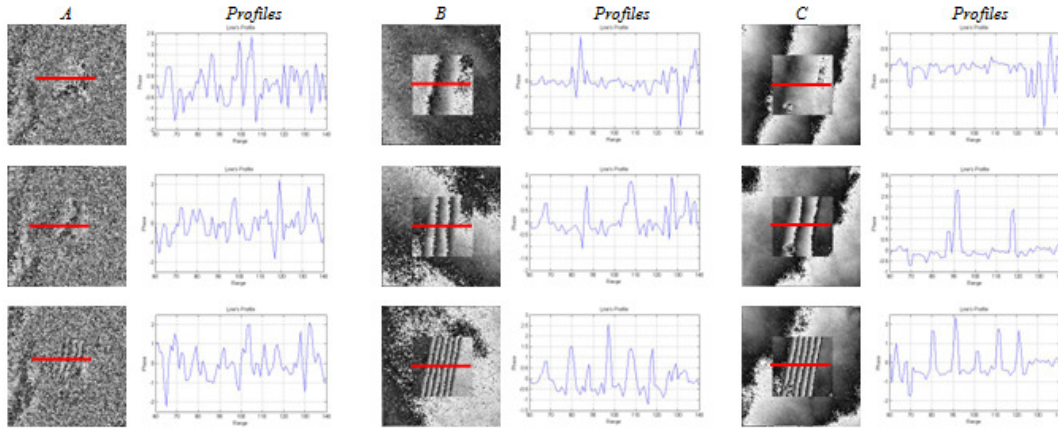


Figure 4. Differentials interferograms and their profiles (red lines), corresponding to simulated deformations SD1 to SD3 with different coherence values (a) $\gamma_A = 0.582355$ , (b) $\gamma_B = 0.679226$  and (c) $\gamma_C = 0.778999$ , for a resolution of 20 m

These results show that the more the coherence is, the best the deformation fringes are distinguished and detected. For low coherence values, no deformation can be detected properly, (e.g., image A). On the other hand, for a same coherence value, the more the deformation gradient is, the more the deformation fringes are distinguished and detected.

Fig. 5 shows the differentials interferograms corresponding to deformations SD1 to SD3 with and without filtering for a 20 m resolution. These results show that the noise spatial filtering makes the fringes contours more readable, thus facilitating the detectability analysis.

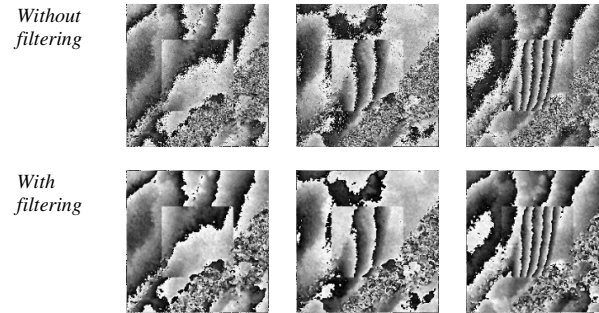


Figure 5. Differentials interferograms corresponding to simulated deformation SD1 to SD3 with a coherence equals to  $\gamma_D = 0.697129$ , for a resolution of 20 m with and without filtering

Regarding the resolution, we have noticed from the results shown in Fig. 6, that the differentials interferograms using a 40 m resolution are more smoothed than those using an 8 m resolution. However, if the deformation is very deep, the number of fringes increases and their width decreases, decreasing the resolution leads to a fringes elimination and an erroneous estimation of the deformation. In the same analysis sense, we have noticed that the simulated images for which the deformation fringes are more tighter and the deformation gradient is big (e.g., SD3), are better observed in the case of interferograms using a 20 m resolution. This result presents a better compromise between the use of a 40 m and 8 m resolutions.

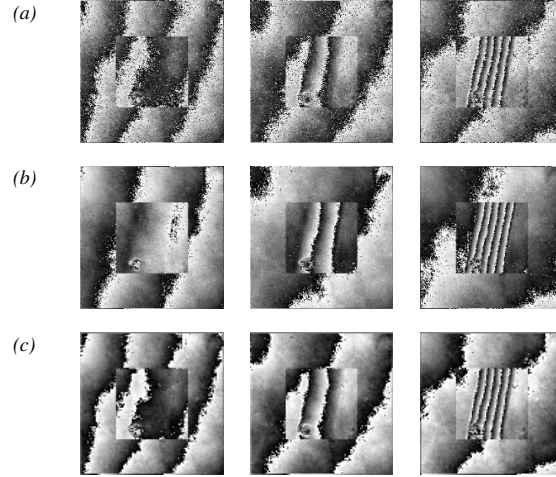


Figure 6. Differentials interferograms corresponding to simulated deformation SD1 to SD3 and a coherence equals to  $\gamma_c = 0.778999$ , with different resolutions: (a) 8 m, (b) 20 m and (c) 40 m

#### 4. SURFACE DEFORMATION GRADIENT MODELLING BY DInSAR

The objective of DInSAR deformation detectability analysis is establishing two deformation gradient equations in function of the coherence (see Fig. 7). The linear curves, obtained from the two equations, allow the delimitation of detectability surface in function of the DInSAR quality product. The intersection, the slopes and the offset of the two curves define the deformation detectability model with respect to types of SAR radar images set to be explored in monitoring and mapping applications of land surfaces and/or Over-surfaces movements.

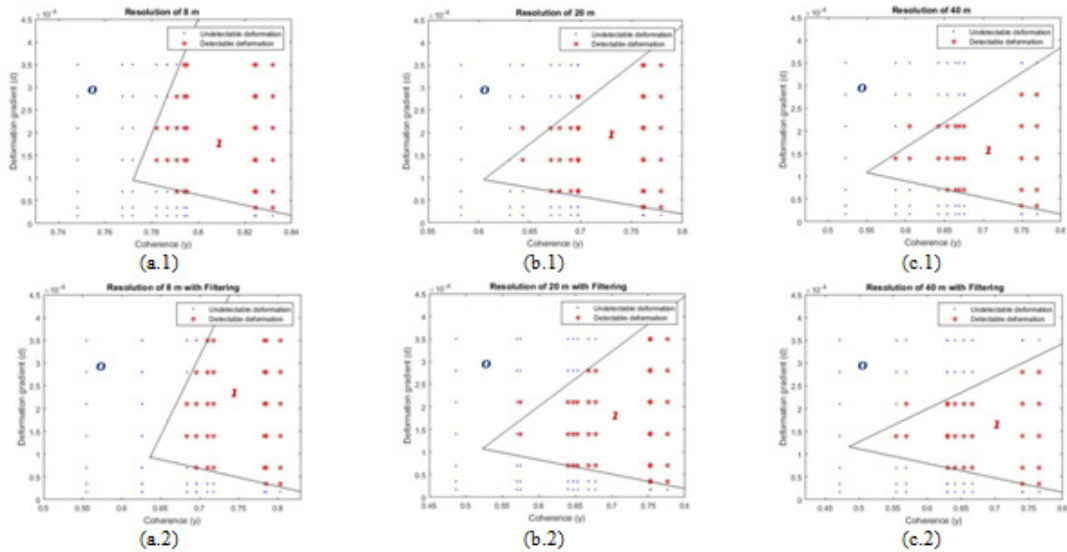


Figure 7. Observations and linear models  $d_{min}(\gamma)$  and  $d_{max}(\gamma)$ , for a resolution (a) 8 m, (b) 20 m and (c) 40 m, (1) without and (2) with spatial filtering

#### 4.1. Model Parameters

This model is defined by two parameters: the coherence and the deformation gradient [6]. The coherence describes the correlation degree between the two SLC images, the modified master image MS and slave image S; it is given by:

$$\gamma = \frac{|\sum_{i=1}^L MS_i \cdot S_i^*|}{\sqrt{\sum_{i=1}^L |MS_i|^2 \sum_{i=1}^L |S_i|^2}} \quad (3)$$

Where:  $L$  is the number of looks corresponding to its appropriate resolution value.

The deformation gradient  $d$  is defined as:

$$d = \frac{DA}{r} \quad (4)$$

Where:  $DA$  is the deformation amplitude and  $r$  is the images resolution according to the range direction.

#### 4.2. Interactive Model Analysis

Each differential interferogram, obtained for a given coherence threshold and a given resolution, associated with the 2D profiles plots are visually analyzed, the decision making on the detectability is related to the fringes limits perception: discernable fringes. On a  $d = f(\gamma)$  graph, we attribute a value '1' if it is discernible, otherwise the value is equal to zero. Fig. 7 represents the deformation gradients ( $d$ ) plots as a function of the coherence ( $\gamma$ ), for the different resolution values (8 m, 20 m and 40 m), with and without the spatial filtering application. By observing the six graphs, the clouds of points forming the decision that equals 1, represented by the red stars can be delimited by two linear curves (one upper bound and another lower one) [6][8]. The minimum and the maximum SDG differ from one resolution to another, and for the same resolution, the spatial filtering makes it different. As the resolution decreases, the SDG upper bound becomes lower; the same happens for the minimum SDG, this means that the  $Max(SDG)$  and the  $Min(SDG)$  vary in the same direction as the resolution.

Linear plots are proposed to approximate the SDG lower and upper boundaries. For this purpose, we obtained a deformation gradient modelling with a resolution of 8 m, 20 m and 40 m respectively, which their equations are the following (with a normalization factor of  $(10^{-4})$ ):

- Without spatial filtering

$$\begin{cases} d_{min} = (9,7504 - 11,4 \gamma) \\ d_{max} = (-97,241 + 127,2 \gamma) \end{cases} \quad (5.a)$$

$$\begin{cases} d_{min} = (3,3064 - 3,89 \gamma) \\ d_{max} = (-9,625 + 17,5 \gamma) \end{cases} \quad (6.a)$$

$$\begin{cases} d_{min} = (3,085 - 3,6496 \gamma) \\ d_{max} = (-4,954 + 10,989 \gamma) \end{cases} \quad (7.a)$$

- With spatial filtering

$$\begin{cases} d_{min} = (3,427 - 3,919 \gamma) \\ d_{max} = (-21,35 + 35 \gamma) \end{cases} \quad (5.b)$$

$$\begin{cases} d_{min} = (2,735 - 3,18 \gamma) \\ d_{max} = (-5,293 + 12,17 \gamma) \end{cases} \quad (6.b)$$

$$\begin{cases} d_{min} = (2,699 - 3,1731 \gamma) \\ d_{max} = (-2,307 + 7,162 \gamma) \end{cases} \quad (7.b)$$

On the other hand, each graph illustrates two areas "1" and "0", which cover the surface where the coherence and the SDG value designate the deformations that can or not be detected by DInSAR. The equation below defines the condition of detectability:

$$\begin{cases} d_{min} \leq d \leq d_{max} \Rightarrow Area "1" \\ d \leq d_{min} \text{ or } d \geq d_{max} \Rightarrow Area "0" \end{cases} \quad (8)$$

This study shows the usefulness of the approach that we propose during the choice of the technique to be implemented for a good detection of the surface deformation. The establishment of the equation (8) will make it possible to decide if a deformation (linear fault) is detectable or not by differential interferometry, taking into account the noise level in the differential interferogram and by calculating only two parameters: the coherence and the gradient, that describe the noise level and the deformation depth respectively.

### 4.3. Discussion

The proposed methodological approach is based on both simulated and real data, the advantage of using simulated images is to be able to control the dimensions of the deformation. On the other hand, introducing the deformation signature into SLC phase images allows reducing the chances of obtaining interferograms with hidden information of the deformation caused by noise. By comparing the deformation image and the resulting differentials interferograms, with the profiles plots, a set of observations has led us to the establishment of a deformation modelling. In practice, this last is useful to evaluate the capacity of the differential interferometry to detect a surface deformation and to monitor its evolution. However, this model has essentially two limitations. In the first place, the detectability decision depends on an interactive method, which makes it very subjective. In addition, the upper and lower boundaries of the model are not certain. Otherwise, this model has been created based on the deformations modeled by simulation. Although the surface deformations are simulated in the form of a brittle surface deformation designating a linear fault, the methodological approach can be generalized to other kinds of deformations. However, in other kind of deformations, it is not practical to calculate the deformation gradient using equations (5) to (7). Therefore, it is more appropriate to adapt the same methodology in order to simulate specific deformations.

## 5. GENERALIZATION

The objective of the generalization is to validate the results obtained in the previous section by checking the reliability of the relationship established between the coherence and the deformation gradient for real data. Since there are no other methods dealing with the same type of deformation and having the same objective and in order to generalize our proposed approach, we used a differential interferogram of the region of Ouargla (Algeria) containing a brittle surface deformation (see Fig. 8).



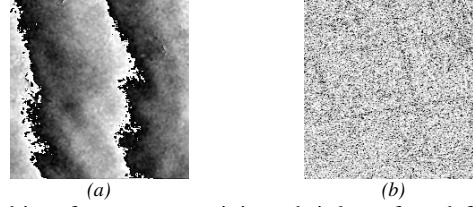


Figure 8. (a) Real differential interferogram containing a brittle surface deformation and (b) its coherence image

This interferogram was generated for a 20 m resolution, we have calculated the deformation gradient, as well as the coherence value. Visually, the deformation is detectable, to evaluate the robustness of this modelling, we assessed the found parameters values in order to determine if the point generated belongs actually to the area “1”. The results obtained are shown in the following table:

Table 2. Calculated parameters.

Gradient	Coherence	$d_{min}$	$d_{max}$
0.000140	0.578806	$0.8944 \times 10^{-4}$	$1.7511 \times 10^{-4}$

From the equation (8), we have noticed that the deformation gradient value fills the following condition:  $d_{min} \leq d \leq d_{max}$ , this implies that this deformation is part of the area “1”, then it is detectable by radar differential interferometry DInSAR.

## 6. CONCLUSIONS

The objective of this study is to show the contribution and the limits of the differential interferometry in altimetric surfaces and sub-surfaces variations detection and estimation. However, these variations altered by diverse degradation sources contributing to the limitations of the DInSAR and which can be highlighted by an interferometric quality indicator which is the coherence. In this sense, we have analyzed and modeled the surface deformation gradient (SDG) by establishing a relationship between the coherence and the deformation gradient for different resolution values. However, an extension of this work will be possible. Indeed, we have only used a visual detectability analysis for the decision in model construction. Other methods based on an automatic analysis can be adopted. In addition, for a reason of simplicity, linear models were used to separate the regions describing detectable and undetectable deformations. Non-linear models can be envisaged.

## REFERENCES

- [1] S.N. Madsen, H. Zebker, & J. Martin, (1993) “Topographic mapping using radar interferometry: processing techniques”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 31, No. 1, pp246-256.
- [2] X.L. Ding, G.X. Liu, Z.W. Li, Z.L. Li, & Y.U. Chen, (2004) “Ground subsidence monitoring Hong Kong with satellite SAR interferometry”, Photogrammetric Engineering and Remote Sensing, Vol. 70, No. 10, pp1151-1156.
- [3] D. Massonnet & K. L. Feigl, (1998) “Radar interferometry and its application to changes in the earth’s surface”, Rev. Geophys., Vol. 36, No. 4, pp441–500.
- [4] H. A. Zebker & J. Villasenor, (1992) “Decorrelation in interferometric radar echoes,” IEEE Transactions on Geoscience and Remote Sensing, Vol. 30, No. 5, pp 950–959.

- [5] X. L. Ding, Z. W. Li, C. Huang & Z.R. Zou, (2007) "Atmospheric effects on repeat-pass InSAR Measurement over SHANGAI Region", J.Atmos. Sol-Terr Phy, Vol. 69, No. 12, pp 1344-1356.
- [6] I. Baran, M. Stewart, &S. Claessens, (2005) "A new functional model for determining minimum and maximum detectable deformation gradient resolved by satellite radar interferometry", IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, No. 4.
- [7] S-H. Yun, H. Zebker, P. Segall, A. Hooper, & M. Poland, (2007) "Interferogram formation in the presence of complex and large deformation", Geophysical Research Letters, Vol. 34.
- [8] Z. W. Li, X. Ding, Z. Jian-Jun, & G. Feng, (2011) "Modeling minimum and maximum detectable deformation gradients of interferometric SAR measurements", International Journal of Applied Earth Observation and Geoinformation, Vol. 13, No. 5, pp 766-777.
- [9] Q. J. Wang, Z. W. Li, Y.N. Du, R. A. Xie, X.Q. Zhang, M. Jiang, & J-J. Zhu, (2014) "Generalized functional model of maximum and minimum detectable deformation gradient for PALSAR interferometry", Trans. Nonferrous Met. Soc. China, Vol. 24,pp 824–832.
- [10] K. Hadj Rabah, F. Hocine, & A. BelhadjAissa, (2017) "Automatic Detection of Surface Deformations by DInSAR", in Proceedings of 6th International Conference on Telecommunication and Remote Sensing.

## AUTHORS

**Karima Hadj-Rabah** received the M.Sc. degree in Telecommunication, networks and multimedia from University of Science and technology HouariBoumedienne(USTHB), Algeria, in 2016. She is currently working toward the Ph.D.degree in synthetic aperture radar (SAR) tomography reconstruction using Very High-resolution data, in laboratory of image processing and radiation, USTHB, under the supervision of Professor AichoucheBelhadj-Aissa.



**Faiza Hocine** Graduated from the University of Sciences and Technology Houari Boumedienne (USTHB), Algeria. PhD in image processing and remote sensing of the same university, in 2015. Currently, she is researcher in Electronics, image processing, remote sensing at USTHB. Her research interests include satellite image processing, SAR interferometry radar.



**SawsenBelhadj-Aissa** received the M.Sc. degree in Telecommunication, networks and multimedia from University of Science and technology HouariBoumedienne (USTHB), Algeria, in 2013. She is currently working toward the Ph.D. degree in SAR Differential Interferometry: 2D / 3D applied to the detection of surface's deformations, in laboratory of image processing and radiation, USTHB, under the supervision of Professor Salah Boughacha.



**AichoucheBelhadj-Aissa** obtained her engineering degree in electronics from National Polytechnic School, Algiers, the master degree and the Doctorate in image processing and remote sensing from the University of the Sciences and Technology Houari Boumedienne, Algiers. Currently, she is university Professor and head of the research team "GIS and integration of geo-referenced data". The main research themes focus on modeling and analysis of textures and forms, fusion and classification of objects, SAR interferometry-polarimetry and GIS.



## **AUTHOR INDEX**

*Adil AbdUlhur AboShana 115*

*Ahmed Gomaa 01*

*Aichouche Belhadj-Aissa 127*

*Akella Amarendra Babu 65*

*Alla Defallah Alrehily 47*

*Amira GHERBOUDJ 33*

*Constantine's Koutsojannis 75*

*Eman Mohamed Nabil Alkholy 75*

*Enoch Agyepong 13*

*Faiza Hocine 127*

*Karima Hadj-Rabah 127*

*Kevin Jones 13*

*Khalid Aljasser 107*

*Mohamed Hassan Haggag 75*

*Mohammad Alshayeb 95, 107*

*Muazzam Ahmed Siddiqui 47*

*Ramadevi Yellasiri 65*

*Sajjad Mahmood 107*

*SawsenBelhadj-Aissa 127*

*Seyed M Buhari 47*

*William J. Buchanan 13*