STUDY OF ALGORITHMS APPLIED TO VOICE SIGNAL ANALYSIS: RECOGNITION OF VOICE PATTERNS USING ARTIFICIAL INTELLIGENCE

Amivaldo Batista dos Santos[1]
Solange da Silva[2]
Pedro José Abrão [3]
Luana Machado dos Santos[4]

ABSTRACT

This study refers to the perspective of voice signal standard recognition using Artificial Intelligence (AI) through Artificial Neural Networks techniques (ANN), using the Mel Frequency Cepstral Coefficients (MFCCs), that extracts voice signal characteristics. The purpose of this paper was recognizing voice signal patterns from the Fire Department of the State of Goiás database. The literature review allows to analyse algorithms that analyse voice signal. Posteriorly, experiments with the same algorithms were done in order to find a pattern of recognition that could identify voice signal characteristics of a phone conversation, in which the facts could be classified as probably true or false. The results of the experiments done with the MFCC algorithm along with AI demonstrated that the extraction of the voice signal characteristics provided by the MFCC is consistent and permit to model a database to study the recognition of voice signal patterns. Thereby, it was possible to identify the user and the attendant's voice in a phone call, as well as identify the characteristics of the female, male and child's voice.
Keywords: Pattern recognition. Voice analysis. MFCC. AI-Mel Frequency.

ESTUDO DE ALGORITMOS APLICADOS À ANÁLISE DE SINAIS DE VOZ: RECONHECIMENTO DE PADRÕES DE VOZ UTILIZANDO INTELIGÊNCIA ARTIFICIAL

RESUMO

Este estudo se refere à perspectiva de reconhecimento padrão de sinal de voz utilizando Inteligência Artificial (IA) por meio de técnicas de Redes Neurais Artificiais (RNA), usando os Coeficientes Cepstrais de Frequência Mel (MFCCs), que extraem características do sinal de voz. O objetivo deste artigo foi reconhecer padrões de sinal de voz a partir do banco de dados do Corpo de Bombeiros do Estado de Goiás. A revisão da literatura permite analisar algoritmos que analisam o sinal de voz. Posteriormente, experimentos com os mesmos algoritmos foram realizados para encontrar um padrão de reconhecimento capaz de identificar características de sinais de voz de uma conversa telefônica, na qual os fatos poderiam ser classificados como provavelmente verdadeiros ou falsos. Os resultados dos experimentos realizados com o algoritmo MFCC juntamente com a IA demonstraram que a extração de características do sinal de voz fornecidas pelo MFCC é consistente e permitiu modelar um banco de dados para estudar o reconhecimento de padrões de sinais de voz. Assim, foi possível identificar o usuário e a voz do atendente em uma chamada telefônica, bem como identificar as características da voz feminina, masculina e infantil.
Palavras-chave: Reconhecimento de padrões. Análise de voz. MFCC. IA-Frequência Mel.

Received August 28, 2023. Approved December 07, 2023

[1] Mestre em Engenharia de Produção e Sistemas - PUC – Goiás. Graduado em Tecnologia em Processamento de Dados pelo Centro Universitário de Goiás - Uni-Anhanguera. Especialista em em Orientação a Objetos e Internet pela Uni Anhanguera. Atualmente professor no Instituto Federal Goiano. E-mail: amivaldo@gmail.com
[2] Doutora em Engenharia Elétrica pela Universidade Federal de Uberlândia. Mestre em Engenharia Elétrica e de Computação pela Universidade Federal de Goiás. Graduada em Ciências - Habilitação em Matemática pela Pontifícia Universidade Católica de Goiás, Especialista em Computação pela PUC Goiás. Atualmente professora na PUC Goiás. E-mail: solansilva.ucg@gmail.com
[3] Doutor em Engenharia Elétrica pela Universidade Federal de Itajubá. Mestre Engenharia Elétrica pela Universidade Federal de Uberlândia. Graduado em Engenharia Elétrica pela Universidade Federal de Uberlândia. Atualmente professor titular do Instituto Federal de Educação, Ciência e Tecnologia de Goiás. E-mail: pedro.abrao@ifg.edu.br
[4] Mestre em Engenharia de Produção e Sistemas pela PUC Goiás, Especialista em Gestão Empresarial com Ênfase em Consultoria pela UNIGOIÁS, Metodologias Ativas e Tecnologias Educacionais. Graduada em Administração pela PUC Goiás. Atualmente professora no Centro Universitário Araguaia. E-mail: luana.santos.adm@hotmail.com

INTRODUCTION

The studies of the pattern recognition and the algorithms implementation with the use of Artificial Intelligence (AI) has made possible a breakthrough in voice recognition. With this development the computational improvements have presented machines that were able to analyse and synthesize the human voice (VERMA, KUMAR, KAUR, 2018).

In the field of psychology, the efforts of researchers to establish a standard for the treatment and recognition of individuals is noticed, mainly in cognitive psychology, which is an area of investigation that can be seen in the publication of Neisser's book (1967). However, the cognitive approach was disclosed by Broadbent, in 1958, in his book Perception and Communication. Since then, the dominant paradigm in the area has been information processing, a model defended by Broadbent. In this line of thought, it is considered that mental processes are comparable to software to be executed by computers that in this case would be the brain. Theories of information processing are based on notions such as: input, representation, computing or processing and outputs (MATLIN, 2004).

In the 20th century, cognitive psychology received a major boost through AI studies, which allows to relate and compare, in a certain way, the human and animal processing of information with electronic processes, such as the computer. As a theory of human behavior, cognitive psychology has emerged as an alternative (ANDERSON, 2004).

Theodoridis and Koutroumbas (2006, p. 58) claim that automatic recognition of emotions is an interdisciplinarity, i.e., a field of research that deals with the algorithmic detection of human affection, such as: anger or sadness, from a variety of other sources, such as speech or facial gestures. Each one is combined with certain advantages and difficulties. Modalities such as: speech, facial gestures or body pose are relatively easy to detect (for example, through a microphone or a camera) and interpret by humans.

Human emotions are laborious to characterize and categorize (ORTONY; CLORE; COLLINS, 1988). The machine learning for the recognition of human emotions is constantly improving. The emotion recognition solutions depend on which emotions we want a machine to recognize and why (YACOUB et al., 2003).

Sentimental or opinion mining analysis is the field of study that analyzes people's attitudes, emotions, feelings and opinions towards entities, such as products, services, organizations, events, topics and the attributes of those collectives. It is a challenging field in Neural Processor Language (NPL), since it deals with several issues of how it is treated through the negation and withdrawal of keywords and can cover many other problems, from the classification in relation to the polarity of an opinion to the process of summarizing the general feeling about something (LIU, 2012).

The algorithms for the signal analysis existing in voice literature were studied, aiming to understand how they were used. This kind of investigation provided a summary of the evidence related to experiments, exploratory studies, among others. That way, this article will serve as a basis for studying algorithms that analyze voice signals, through the application of explicit and systematic methods of searching, critical appreciation and synthesis of the selected information, enabling a possible voice recognition pattern. Thus, facts are indicated in a telephone conversation which can probably be classified as true or false. The voice recognition pattern in a telephone conversation may assist the telephone answering service of emergency services (for example: Fire Department) that daily receive false information on telephone calls.

According to the context, this article aims to answer the following research question: Can the MFCC algorithm, together with AI techniques, assist in voice pattern recognition in a telephone call?

This article is organized as follows: Section 1 provides the introduction, in which the

purpose of the article is addressed. Section 2 presents the theoretical reference with concepts, definitions and related projects. Section 3 presents the methodology used in the analysis of this project. Section 4 concludes the studies carried out.

Theoretical Reference

Countless social networks express emotions and feelings about companies and their services; it is possible to foresee several stressful situations with customers and, with that, search for better service in order to achieve continuous improvement through feedbacks and possible preventive actions; internally analyzing corporate service. The attendant is usually exposed to training and various adjustments for a good service, however it is impossible to analyze, at that moment, all possible deviations from conduct and reckless cases in gauged words (LIU, 2012).

In a Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictive Cepstral Coefficients (LPCC), voice resources are often used. Consecutively, the Gaussian Mixture Model (GMM) is normally used as a classifier for the recognition of the person who speaks (speaker) (AWAIS et al., 2018).

YU et al. (2001), used Support Vector Machines (SVMs) to detect emotions. They built classifiers for four emotions: anger, joy, sadness and neutral. Since SVMs are binary classifiers, their recognizers worked on detecting one emotion versus the rest. An average accuracy of 73% has been reported.

For the abstraction algorithm, speech is reduced in up to 13 different sub-bands, revealed by MFCC filtering, followed by Discrete Fourier Transform (DFT). MFCC filtering can simulate human sound related features and improve individual speech intelligibility. Although DFT cannot efficiently examine the non-stationary, in addition to the non-linear voice signal. The speech is broken into 9 sub-bands, separated by the Discrete Wavelet Transform (DWT). That way the voice signal becomes energy transmission, owning 9 energy probabilities and drifts to make a vector of characteristics, extracted through these sub-bands. In order to solve the DFT problem, the DWT can examine the conversation successfully. In addition, it can react with MFCC filtering, considering the resource related to human sound. However, the sudden changes that occur in resources related to human sound cannot be recognized by the distribution of energy. The MFCC represents one of the main descriptives forms of signal (AWAIS et al., 2018).

According to Mohamed (2012), in automatic speech recognition, he lists some tools that work with machine learning with a supervised and unsupervised approach, such as:

- **SenticNet**, which makes up a set of tools and techniques for analyzing feelings that combine reasoning with common sense, psychology, linguistics and machine learning;
- **SentiWord**, based on the Mexico WordNet dictionary, which groups adjectives, verbs and other grammatical classes into sets (synset) in a semi-supervised approach;
- **Lexicon Sentiment1401**, also known as TwitterSentiment, allows you to discover feelings related to a brand, product or topics on Twitter;
- **Valence AwareDictionary for SentimentReasoning (VADER)**, sentiment analysis tool, developed to evaluate messages in the context of Twitter and other online social media that does not require training.

YACOUB et al. (2003), considers that the performance of a voice classifier depends a lot on the quality of the database used for training and testing and its similarity with real samples (generalization). Speech data used to recognize voice tests can be grouped into three categories, depending on how the speech signal was recorded. The first method uses actors to record sayings. Each statement is spoken with various faked emotions. The actors are given time to imagine themselves in a specific situation before speaking. The second method, called OfOz Assistant, uses a program that interacts with the actor and takes him into a specific emotional situation and

then records his responses.

The third method, which is harder to obtain, is the recording of real-world statements that express emotions. Automatic speech recognition consists of two parts. The first part is training part in which the entire speaker database is created and the other part is testing the one on which occurs speech recognition. Different phases of speech recognition are: Feature Extraction (speech analyzer) and Matching process (Speech Classifier). Feature Extraction is necessary because there is not much variability in the digital waveform, so it reduces variability. Extraction of characteristics take out resources that can be analyzed in loudspeaker. There are different techniques of extracting resources presented as Linear Predictive Coding (LPC), Perceptual Linear Coding (PLC) and MFCC, etc. (VERMA, KUMAR, KAUR, 2018).

MFCC represents the power spectrum for the speech signal on the basis of the transformation of the voice signal. On the MFCC frequency scale, the linear frequency spacing is less than 1000Hz and for a log it is greater than 1000Hz. In the process of extracting characteristics, continuous speech is introduced for windowing. After that, a voice signal, which is in a continuous form, is converted into window frames. These frames are then transmitted to the Fourier Transform process, which transforms window frames into a spectrum, generating analyzes of the voice signal (VERMA, KUMAR, KAUR, 2018).

The EMOSpeech tool is being widely used to analyze speech, allowing call center companies to analyze recorded calls. It is based on a model that recognizes different emotions in a spectrum of acoustic emotions and this contributes to identify the emotion using some properties of the voice. Based on several samples, the program is able to detect anger, joy, sadness, tranquility. In addition, the company provides initials (API) for developers (VERMA, KUMAR, KAUR, 2018).

METHODOLOGY

This research consisted of a theoretical study, raising the concepts and definitions about the theme and describing the tools to be used and a practical study, looking for the related literature works, through book reviews, theses, essays and regular articles of the area. The experiments were carried out using thes SenticNet, SentiWord, Lexicon Sentiment1401 and Vader tools.

As for the procedures, this research can be classified as a case study and experimental study. The case study is used in both biomedical and social sciences, which consist of a detailed study of a research object, which allows to have knowledge about the subject (GIL, 2002).

As for the technical procedures, the formulation of the problem, the execution planning, data collection and data analysis are accomplished. In the Case Study, the focus is the collection research of a specific real activity of the Fire Department to identify patterns in the phone calls voice analysis received by them. The experimental research presents the aspect of manipulation of reality by the researcher (MARTINS; MIGUEL, 2012). Thus, it can determine an object of study and there are variables selected that are executed on that target, they are registered as forms of manipulation and observations of the effects of variables on that object (GIL, 2010). It was necessary to have experiments done in order to identify false calls risks, using mechanisms for deepening customer service, observing the effects produced.

After analyzing the literature, a test was performed with the MEL algorithm to do the voice analysis. The MatLab software version 2018B implements this algorithm in the Simulink module. The MFCC technique was chosen because it has better precision in extracting the voice signal characteristics.

It was verified that the analysis performed by the MFCC brings a total of 14 Cepstral coefficients that are extracted from the sub-bands to prepare a vector of characteristics along with the pitch - the word pitch can be understood as the frequency measured in an audio signal.

And through the pitch it is possible to classify the voice signal in: male voice; female and child voice.

The Pitch coefficient after processing in MatLab, results in the frequency of the voice in Hertz (Hz), allowing to analyze whether it is a male, female or child voice. The pitch is also called the fundamental frequency and represents the frequency of vibration of the vocal cords. Values considered low, which are between 85 and 180 Hz, are often male, considering an adult in normal speech. The frequency for female voice can vary from 165 to 255 Hz. For children and mostly babies, the values are above 300 Hz. The pitch is calculated from the standard deviation of the samples from the first MFCC coefficient.

The intermediate layer is composed of 10 neurons that will learn the network; this number was parameterized according to tests that were carried out and it was verified that this is the number necessary to have a good training of the Artificial Neural Network (ANN) presented.

Availability of data and materials

The data supporting the investigations of this study were available at the Military Fire Department of the State of Goiás, but it's worth noting that there is a restriction on the availability of this data, which was used under license for the current study and therefore is not accessible to the public. However, the data has been made available upon reasonable request and with permission from the Military Fire Department of the State of Goiás.

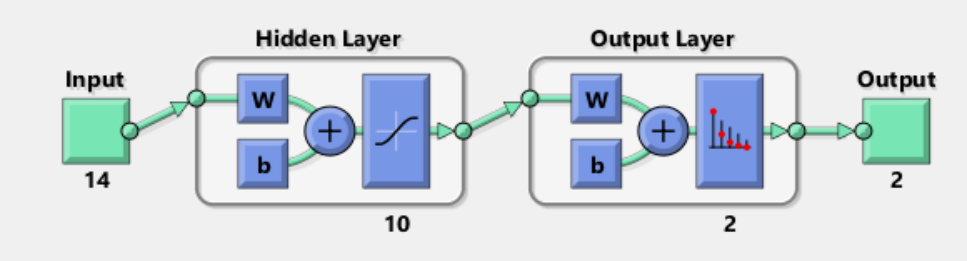
RESULTS AND DISCUSSION

Through the analysis of the database provided by the Fire Department of the State of Goiás, it was possible to verify the learning of the initials ANN, using the nnstart module of MatLab version 2018- B.

From 498 samples for attendants recognition, ANN identified 496 correctly, and only two samples were incorrectly identified. For the 498 samples for recognition, all users (100%) were correctly identified. It is concluded that the ANN had a 99.8% efficiency in training for user and attendant identification.

In the experiments, a matrix of 996 samples (498 of users and 498 of attendants) was used, with 14 MFCC coefficients, which permitted to generate an ANN that to identify an audio file of a phone call, both the user's voice and the attendant's voice. In addition, it also identifies whether it is a male, female or child voice. An ANN was generated and it allowed to identify, in an audio file, both user and attendant, as shown in Picture 1.

Picture 1 – Artificial Neural Network instruction representation for user and attendant identification.



Source: Elaborated by the authors (2020).

The RNA model above consists of 14 inputs, which are the coefficients extracted from the voice with the MFCC algorithm and are constant in the input data matrix which has 996 voice samples, being classified into two files, one for the attendant and the other for the user. Thereby,

only five seconds of all data were extracted, and consequently, the MFCC provided an archive of 498 copies of attendants and users.

The intermediate layer is composed of 10 neurons that will make the learning of the network; this number was parameterized according to tests that were done, and it was verified that this is the right necessary to have a good training of the ANN presented, because the amount of neurons is parameterized empirically, observing the result of learning the neural network.

The output layer has two possible binary responses, which are: 10 to identify that the voice is from an attendant, and 01 to identify that the voice is from a user.

To identify fake calls it is necessary to classify the voice of the user and the attendant, as the user's voice will be subjected to a voice pattern recognition so that the research question can be answered.

Using the data set developed for this work, this step aims to obtain the voice distinction between the attendant and the user. This is done through an Artificial Neural Network and consisted of 14 entries, representing the coefficients extracted from the voice with the MFCC algorithm and are constant in the input data matrix which has 996 copies of voices, 498 of attendants and 498 of users. The middle layer was composed of 10 neurons that learned the network. This number was parameterized in an empirical way and the tests were satisfactory. As it had been considered empirical, ANN was also trained with only two neurons.

Table 1 shows the result of the confusing matrix of ANN training with 10 neurons, 498 copies of attendant voices and 498 copies with user voices, the first column identifies the phase of ANN training, the second column shows the amount of attendant voice copies, and the third column the number of user voice copies. For ANN training a percentage of 70% was used for training, and for validation 15%, and the test phase 15%.

Table 1 – Result of the ANN training of 996 copies with 10 neurons.

Confusing Matrix	**Nº Copy Attendant**	**Nº Copy User**
Training (correct)	353 (50,6%)	345 (49,4%)
Training (incorrect)	0 (0,0%)	0 (0,0%)
Validation (correct)	67(45,0%)	82 (55,0%)
Validation (incorrect)	0 (0,0%)	0 (0,0%)
Test (correct)	78 (52,3%)	71 (47,7%)
Test (incorrect)	0 (0,0%)	0 (0,0%)
Total ANN (correct)	498 (50,0%)	498 (50,0%)
Total ANN (incorrect)	0 (0,0%)	0 (0,0%)
ANN correct rate	**100%**	**100%**

Source: Elaborated by the author (2020)

Table 1 shows that ANN identified all copies of both the user and the attendant without any error. Table 2 shows the ANN training with the same sample as in Table 1, but using 2 neurons.

Table 2 – Result of the ANN training of 996 copies with 2 neurons

Confusing Matrix	**Nº Copy Attendant**	**Nº Copy User**
Training (correct)	353 (50,6%)	345 (49,4%)
Training (incorrect)	0 (0,0%)	0 (0,0%)
Validation (correct)	72(48.3%)	77 (51,7%)
Validation (incorrect)	0 (0,0%)	0 (0,0%)
Test (correct)	73 (49,0%)	76 (51,0%)
Test (incorrect)	0 (0,0%)	0 (0,0%)

Total ANN (correct)	498 (50,0%)	498 (50,0%)
Total ANN (incorrect)	0 (0,0%)	0 (0,0%)
ANN correct rate	**100%**	**100%**

Source: Elaborated by the author (2020)

Table 2 shows that the ANN identified all copies of both user and attendant without any error, only the number of copies of the validation and test that were different from the training with 10 neurons, however the percentage of correct identification of the samples remained with no mistakes.

Noticing the two trainings it is possible to conclude that for this sample of data the training with 2 neurons is just enough for satisfactory results.

The results of the sample submitted to the identification of the attendant and user voice are shown in Table 3, the sample has 498 copies of the attendant voice and 498 copies of the user voice.

Table 3 – Confusion Matrix of 996 copies – Identification of attendant and user

Confusing Matrix	**Nº Copy Attendant**	**Nº Copy User**
ANN correct rate	493 (49,5%)	494 (49,6%)
ANN incorrect rate	5 (0,5%)	4 (0,4%)

Source: Elaborated by the author (2020).

The results in Table 3 show that 493 samples of the 498 sampled to classify attendant are correctly classified. So the error rate is 0.5%. From 498 copies for user classification, 494 were classified correctly and only 0.4% were classified incorrectly. It is concluded that the ANN had an efficiency of 99.1% (correct answers) in the training for user and attendant identification.

A sample with 498 copies of attendant was submitted to ANN and the result is shown in Table 4.

Table 4 – Confusion Matrix of 498 copies – Identification of the attendant

Confusing Matrix	**Nº Copy Attendant**	**Nº Copy User**
ANN correct rate	495 (99,4%)	0 (0,0%)
ANN incorrect rate	3 (0,6%)	0 (0,0%)

Source: Elaborated by the author (2020).

It is noted in Table 4 that the result for identifying the attendant was 99.4% correct. Only 3 copies were not properly identified. As there were only entries of examples of attendants, the confusion matrix shows zeroed results for user identification, that way, the efficiency of the ANN is verified. A classification was also performed with a sample that contains only copies of the user's voice, the results are shown in Table 5.

Table 5 – Confusion Matrix of 498 copies – Identification of the user

Confusing Matrix	**Nº Copy Attendant**	**Nº Copy User**
ANN correct rate	0 (0,0%)	497 (99,8%)
ANN incorrect rate	0 (0,0%)	1 (0,2%)

Source: Elaborated by the author (2020).

Table 5 shows that the result for user identification was 99.8% correct. Only 1 copy was not properly identified. The result for attendant identification is zeroed (0.0%), as there were only entries from users' samples.

FINAL CONSIDERATIONS

This article aimed to confirm whether the MFCC algorithms, along with AI techniques, could assist in the recognition of voice patterns of a phone call, with the target of study to identify fake calls "prank calls" from a phone call. A database of the Fire Department of the State of Goiás was used, using algorithms to extract voice signal characteristics along with AI.

Studies have shown that the extraction of voice characteristics processed in a neural network allows the person to identify who are the attendant and the user. The database allowed them to be presented during the experiments, since the training with the ANN was satisfactory, as the tests carried out showed an assertion of more than 98% in the recognition of patterns.

The results of the experiments carried out with the MFCC algorithm along with AI, demonstrated that the extraction of the characteristics of the voice signal provided by the MFCC is consistent and allow to model a database to study the recognition of voice signal patterns.

Thus, it was possible to identify the voice of the user and the attendant in a telephone call, as well as to identify the characteristics of the female, male and child voice. And from the analysis of the ANN it was concluded that the extractions of the voice characteristics provided by the MFCC are consistent and allow the author to explore a database and, to study voice pattern recognition, in order to direct new researches that can work subjects that relative to the voice signal characteristics.

REFERENCES

AWAIS, Ahmed et al. Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing. In: **2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)**. IEEE, 2018. p. 271-276.

CAUCHICK MIGUEL, Paulo Augusto et al. Metodologia de pesquisa em engenharia de produção e gestão de operações. **Rio de Janeiro: Elzevir**, 2ª Edição, 2012.

GIL, Antonio Carlos et al. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 5ª Edição, 2010.

JOHN R.(PSYCHOLOGE) ANDERSON. **Psicologia cognitiva e suas implicações experimentais**. LTC Ed., 2004.

LIU, Bing. **Sentiment analysis and opinion mining**. Morgan and Claypool Publishers, 2012.

MATLIN, Margareth. **Psicologia cognitiva**. 5º Edição, Rio de Janeiro: LTC, 2004.

MOHAMED, Abdel-rahman; DAHL, George E.; HINTON, Geoffrey. Acoustic modeling using deep belief networks. **IEEE transactions on audio, speech, and language processing**, v. 20, n. 1, p. 14-22, 2012.

ORTONY, Andrew; CLORE, Gerald L.; COLLINS, Allan. The cognitive structure of emotions. 1988.

THEODORIDIS, S. KOUTROUMBAS, K. **Pattern Recognition**, 3ª Edição, São Paulo: Elsevier, 2006.

YACOUB, Sherif M. et al. Recognition of emotions in interactive voice response systems. In: **Interspeech**. 2003.

YU, Feng et al. Emotion detection from speech to enrich multimedia content. In: **Pacific-Rim Conference on Multimedia**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 550-557.

VERMA, Amit; KUMAR, Amit; KAUR, Iqbaldeep. Automatic speech recognition using Mel-

frequency cepstrum coefficient (MFCC) and vector quantization (VQ) techniques for continuous speech. **International Journal of Advanced and Applied Sciences**, v. 5, n. 4, p. 73-78, 2018.